Advanced mathematical statistics August 30, 2025

Part III. Theory for M-estimation and MLE

Instructor: Lin Liu

1 A summary of minimax hypothesis testing and outlook

Usually the proof goes according to the following steps:

1. Ansatz a worst-case perturbation over the null hypothesis that belongs to the alternative: not too far from the null but far enough to be in the alternative;
2. Compute f-divergence between the product measure under the null and the product measure of the worst-case instance to show it is $O(1)$ so difficult to separate information theoretically;
3. At step 2, you have completed the lower bound proof; for upper bound, simply provide a test statistic whose testing risk equals the lower bound.

But note that in reality, one often has a natural test statistic, the rate of which looks "right". Then one proceeds to prove if a matching lower bound exists.

In later lectures, we will revisit hypothesis testing problems in two occasions:

1. Deriving estimation lower bound via hypothesis testing problems.
2. What we have done so far is about deriving the information theoretical limit of hypothesis testing problems: Neyman-Pearson lemma tells us for single vs. single test, likelihood ratio is the optimal test; In general, we can use f-divergence to quantify the "distance" between H_0 and H_a, which further implies if or not the optimal test can succeed in distinguishing between H_0 and H_a. In either case, the search space of all tests is any measurable function of the observed data, with no restriction on the computational hardness of the search space.

   Low-degree polynomial conjecture: in Sam Hopkins' PhD thesis [**?**], he made the following important conjecture

   > **Conjecture 1.** If a statistical problem cannot be solved using statistics based on low-degree polynomials, then the statistical problem is algorithmically/computationally hard.

   This conjecture has been refuted by [**?**], but [**?**] posed a modified conjecture dealing with the counter-examples. In general, people believe this conjecture to be reasonable.

2 Cramér-Rao lower bound and van Trees inequality

Neyman-Pearson is the most elementary lower bound in hypothesis testing. The counterpart in estimation is Cramér-Rao lower bound. It says the following:

Theorem 2 (Cramér-Rao lower bound)**.** *In a statistical model parameterized by θ, $\widehat{\theta}$ is an unbiased estimator of θ based on data $\boldsymbol{X}$, $I(\theta)$ is the corresponding Fisher information. Then*

$$\mathsf{var}_\theta(\widehat{\theta}) \succeq I(\theta)^{-1}. \tag{1}$$

where $I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \ell(\boldsymbol{X};\theta)}{\partial \theta}\right)^{\otimes 2}\right]$ where $\ell(\boldsymbol{X};\theta)$ is the log-likelihood function of parameter θ with data $\boldsymbol{X}$.

Proof. We prove the theorem for the scalar case with n i.i.d. data so $\boldsymbol{X} = (X_1, \cdots, X_n)^\top$ to help you understand all the concepts related to MLE better. Denote the score function for one data as

$$S(X;\theta) = \frac{\partial \ell(X;\theta)}{\partial \theta} = \frac{\partial \log f_X(X;\theta)}{\partial \theta}.$$

Obviously $\mathbb{E}_\theta(S(X;\theta)) = 0$ and $\sum_{i=1}^n S(X_i;\theta)$ is the score for all n data. Then

$$\begin{aligned}
\mathsf{cov}_\theta\left(\widehat{\theta}, \sum_{i=1}^n S(X_i;\theta)\right) &= \mathbb{E}_\theta(\widehat{\theta}\sum_{i=1}^n S(X_i;\theta)) \\
&= \mathbb{E}_\theta\left(\widehat{\theta}\sum_{i=1}^n \frac{1}{f_{X_i}(X_i;\theta)}\frac{\partial f_{X_i}(X_i;\theta)}{\partial \theta}\right) \\
&= \int \widehat{\theta}(x_1, \cdots, x_n)\sum_{i=1}^n \left(\frac{\partial f_{X_i}(x_i;\theta)}{\partial \theta}\prod_{j\neq i} f_{X_j}(x_j;\theta)\right)\mathrm{d}(x_1, \cdots, x_n) \\
&= \int \widehat{\theta}(x_1, \cdots, x_n)\frac{\partial}{\partial \theta}\prod_{i=1}^n f_{X_i}(x_i;\theta)\mathrm{d}(x_1, \cdots, x_n) \\
&= \frac{\partial}{\partial \theta}\int \widehat{\theta}(x_1, \cdots, x_n)\prod_{i=1}^n f_{X_i}(x_i;\theta)\mathrm{d}(x_1, \cdots, x_n) \\
&= 1.
\end{aligned}$$

where the last line is due to the unbiasedness of $\widehat{\theta}$. Lastly, we use Cauchy-Schwarz inequality to bound 1:

$$\begin{aligned}
1 &\leq \mathsf{var}_\theta(\widehat{\theta})\mathsf{var}_\theta\left(\sum_{i=1}^n S(X_i;\theta)\right) \\
\Rightarrow \mathsf{var}_\theta(\widehat{\theta}) &\geq \frac{1}{\mathsf{var}_\theta\left(\sum_{i=1}^n S(X_i;\theta)\right)} = \frac{1}{I_n(\theta)}.
\end{aligned}$$

□

How about Bayesian analog? Just marginalizing over the above inequality:

$$\mathbb{E}_{\theta\sim\Pi}\left[\mathsf{var}_\theta(\widehat{\theta})\right] \succeq \mathbb{E}_{\theta\sim\Pi}\left[I(\theta)^{-1}\right]. \tag{2}$$

Here comes van Trees inequality, which does not need $\widehat{\theta}$ to be unbiased if we introduce a prior.

Theorem 3 (van Trees)**.** *In a statistical model parameterized by θ, $\widehat{\theta}$ is any estimator of θ based on data X, $I(\theta)$ is the Fisher information of X. For a suitably chosen (prior) probability measure Π with density π on Θ, we have*

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim\mathbb{P}_\theta}(\widehat{\theta}-\theta)^2\right] \geq \left\{\mathbb{E}_{\theta\sim\pi}\left[I(\theta)\right] + I(\pi)\right\}^{-1} \tag{3}$$

where $I(\pi) = \int_{\theta\in\Theta}\left(\frac{\partial \log \pi(\theta)}{\partial\theta}\right)^{\otimes 2}\pi(\theta)\mathrm{d}\theta$ is the Fisher information of the prior.

Proof. The proof is elementary. We introduce a perturbed joint distribution of the likelihood and prior:

$$\pi_h(x,\theta) = p(x|\theta+h)\pi(\theta+h) \tag{4}$$

for h small. Note that for this perturbation to be well-defined, we need to choose π to be smooth with compact support. With this new perturbed measure, we can show

$$\begin{aligned}\mathbb{E}_h(\widehat{\theta}-\theta) &= \int_\theta\int_x (\widehat{\theta}(x)-\theta)p(x|\theta+h)\pi(\theta+h)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\theta'}\int_x (\widehat{\theta}(x)-\theta'+h)p(x|\theta')\pi(\theta')\mathrm{d}x\mathrm{d}\theta' \\ &= \mathbb{E}_0(\widehat{\theta}-\theta)+h.\end{aligned} \tag{5}$$

Define the first order difference

$$D_h(x,\theta) = \frac{\pi_h(x,\theta)-\pi_0(x,\theta)}{\pi_0(x,\theta)}. \tag{6}$$

Obviously

$$\mathbb{E}_0 D_h(x,\theta) = \int_\theta\int_x\left(\frac{\pi_h(x,\theta)}{\pi_0(x,\theta)}-1\right)\pi_0(x,\theta)\mathrm{d}x\mathrm{d}\theta = 0$$

which is essentially the score. Now, invoking (5), we have

$$\mathbb{E}_0 D_h(x,\theta)(\widehat{\theta}-\theta) = \mathbb{E}_h(\widehat{\theta}-\theta) - \mathbb{E}_0(\widehat{\theta}-\theta) = h.$$

Similar to the proof of Cramér-Rao lower bound, we use Cauchy-Schwarz inequality

$$\begin{aligned}&\mathbb{E}_0 D_h(x,\theta)^2\mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\ \Leftrightarrow\ &\int\frac{(\pi_h(x,\theta)-\pi_0(x,\theta))^2}{\pi_0(x,\theta)}\mathrm{d}x\mathrm{d}\theta\mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\ \Leftrightarrow\ &\left(\int\frac{\pi_h(x,\theta)^2}{\pi_0(x,\theta)}\mathrm{d}x\mathrm{d}\theta-1\right)\mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\ \Leftrightarrow\ &\left(\int\frac{\pi(\theta+h)^2}{\pi(\theta)}\frac{p(x|\theta+h)^2}{p(x|\theta)}\mathrm{d}x\mathrm{d}\theta-1\right)\mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\ \overset{\approx}{\Leftrightarrow}\ &\left(\int\frac{\pi(\theta+h)^2}{\pi(\theta)}\left(1+h^2I(\theta)\right)\mathrm{d}x\mathrm{d}\theta-1\right)\mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2\end{aligned}$$

$$
\begin{aligned}
&\overset{\approx}{\Leftrightarrow} \left(\int \frac{\pi(\theta+h)^2}{\pi(\theta)} \mathrm{d}x\mathrm{d}\theta - 1 + h^2 \int I(\theta)\pi(\theta)\mathrm{d}\theta \right) \mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\
&\Leftrightarrow \left(1 + h^2 I(\pi) - 1 + h^2 \int I(\theta)\pi(\theta)\mathrm{d}\theta \right) \mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2 \\
&\Leftrightarrow \left(h^2 I(\pi) + h^2 \int I(\theta)\pi(\theta)\mathrm{d}\theta \right) \mathbb{E}_0(\widehat{\theta}-\theta)^2 \geq h^2.
\end{aligned}
$$

□

van Trees inequality has been shown to be related to the famous Local Asymptotic Minimax (LAM) theorem by David Pollard and colleagues [**?**], which we will discuss briefly at the end of this chapter.

3 M-estimation theory

M/Z-estimation is an enormously general estimation framework. Any estimation procedure that is itself solving an optimization problem is M-estimation:

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{M}(X_1, \cdots, X_n; \theta)$$

for some objective function $\mathcal{M}$. If $\mathcal{M}$ is differentiable and the Hessian is negative semidefinite, then the optimization problem reduces to an estimating equation, which is Z-estimation.

When $\mathcal{M}$ is the data (pseudo/profile/partial/...) log-likelihood, $\widehat{\theta}$ is the (pseudo/profile/partial/...) MLE; when $-\mathcal{M}$ is the empirical risk (very common in machine learning), $\widehat{\theta}$ is the empirical risk minimization (ERM) estimator; when $-\mathcal{M}$ is empirical risk plus some regularization terms, then $\widehat{\theta}$ is the penalized ERM estimator... In the most general form, $\mathcal{M}$ can be strongly convex, convex, non-convex, smooth, or even non-smooth, and does not have to be the form of sum of i.i.d.

Now let us focus on the following type of M-estimation, which includes MLE as a special case. Denote $m(O, \theta)$ a function of some random vector $O \in \mathbb{R}^d$ that depends on unknown parameter θ, $M(\theta) \coloneqq \mathbb{P}[m(O, \theta)]$ its population expectation, and $\mathbb{M}_n(\theta) \coloneqq \mathbb{P}_n[m(O, \theta)]$. Here we introduce in this lecture the following two operators: $\mathbb{P}[\cdot]$ the population expectation operator with respect to the true law, and $\mathbb{P}_n[\cdot]$ the empirical expectation operator.

Definition 4 (Empirical objective function maximization)**.** The true parameter θ_0 is the solution to the following population maximization problem:

$$\theta_0 = \arg\max_{\theta \in \Theta} M(\theta). \tag{7}$$

Then the corresponding M-estimator $\widehat{\theta}_n$ is the solution to the following empirical maximization problem:

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \mathbb{M}_n(\theta). \tag{8}$$

In this lecture, we consider mostly the case that $\Theta = \mathbb{R}^d$. When this is not the case, we are dealing with constrained M-estimation, which under certain regularity conditions is equivalent to penalized M-estimation by duality.

We consider three different layers of statistical properties of $\widehat{\theta}_n$:

1. When $\widehat{\theta}_n$ is a consistent estimator of θ_0?
2. What is the speed at which $\widehat{\theta}_n$ converges to θ_0 in $\|\cdot\|$-norm, as n increases?
3. Is normal a good approximation to the distribution of $\widehat{\theta}_n$ as n grows?

3.1 Consistency of $\widehat{\theta}_n$

Formally, we are trying to show $\|\widehat{\theta}_n - \theta_0\| \to 0$ in $\mathbb{P}_{\theta_0}$-probability. Usually, the intuition goes as follows under the "well-posedness premise"

> there is no "spurious optima" θ', which is a point different from θ_0 but $M(\theta_0) \approx M(\theta')$ for the population optimization program (7), i.e. $M(\theta_0) > \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta')$

$$\|\widehat{\theta}_n - \theta_0\| > \delta \implies M(\theta_0) - M(\widehat{\theta}_n) > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta'). \tag{9}$$

Then by the "zeroth-order" condition of the empirical optimization program (8)

$$\mathbb{M}_n(\widehat{\theta}_n) \geq \mathbb{M}_n(\theta_0), \tag{10}$$

combined with the well-posedness of the problem in (9), we have

$$\begin{aligned}
&[M(\theta_0) - M(\widehat{\theta}_n)] + [\mathbb{M}_n(\widehat{\theta}_n) - \mathbb{M}_n(\theta_0)] > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta') \\
\Leftrightarrow\ &[M - \mathbb{M}_n](\theta_0) - [M - \mathbb{M}_n](\widehat{\theta}_n) > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta') \\
\Leftrightarrow\ &[M - \mathbb{M}_n](\theta_0 - \widehat{\theta}_n) > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta') \\
\Leftarrow\ &o_{\mathbb{P}_\theta}(1) + \sup_{\theta\in\Theta}[M - \mathbb{M}_n](\theta) > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta').
\end{aligned} \tag{11}$$

Thus we ask under what condition of M, we have

$$2\sup_{\theta\in\Theta}[M - \mathbb{M}_n](\theta) > M(\theta_0) - \sup_{\theta:\|\theta_0-\theta'\|>\delta} M(\theta')$$

with negligible probability.

Now we observe the LHS of the above display is

$$\sup_{\theta\in\Theta}[M - \mathbb{M}_n](\theta) \tag{12}$$

which is a stochastic process indexed by θ, and the randomness of this stochastic process is a result of the empirical measure in $\mathbb{M}_n = \frac{1}{n}\sum_{i=1}^n m(O_i, \theta)$. As a consequence, we call (12) the "empirical process" term, which was the central object of study between 1970's and 1990's for probabilists. To my best knowledge, this field was started by Richard Dudley (MIT, passed in 2020), then followed by Evarist Giné (University of Connecticut, passed in 2015), Michel Talagrand[1], Jon A Wellner, Roman Vershynin, Vladimir Koltchinskii, Aad van der Vaart, and younger generations like Richard Nickl and Ramon van Handel. Now the focus of probability has shifted to statistical physics and their connection to computational complexity theory.

[1]For students who are ambitious enough to become probabilists, you should learn more about Michel Talagrand, who has not only made seminal contributions to empirical processes and hence learning theory, but also to the intersection between statistical physics and computational complexity regarding Parisi formula, Ising models and spin glass theory.

3.2 Glivenko-Cantelli (GC) class

An underlying philosophy of empirical process theory is to find the correct complexity measure of $\mathcal{F}$ that suffices to guarantee certain desired properties to happen.

There are in general two types of complexity measures of a function class/a parameter space $\mathcal{F}$:

(1) Metric entropy [**?**] based on covering number or bracketing number;

(2) Combinatorial dimension [**?**] (simplest example: Vapnik-Chervonenkis dimension [**?**]).

In fact, as we will show later, combinatorial dimension is often used to calculate metric entropy.

We start with the following now-standard symmetrization trick. First create independent copies $Y_1, \cdots, Y_n$ of $X_1, \cdots, X_n$ and independent Rademacher random signs $\varepsilon_1, \cdots, \varepsilon_n \sim \mathsf{Rad}(1/2)$. Then

$$\begin{aligned}
\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} &= \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \\
&= \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_Y[f(Y_i) \right| \\
&\le \mathbb{E}_Y \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i) \right| \\
&\le \mathbb{E}_\varepsilon \mathbb{E}_Y \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \{f(X_i) - f(Y_i)\} \right| \\
\mathbb{E}_X \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} &\le 2\mathbb{E}_X \mathbb{E}_\varepsilon \underbrace{\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right|}_{=:\mathsf{Rad}_n(\mathcal{F})} \text{ by Triangle inequality}
\end{aligned}$$

where $\mathsf{Rad}_n(\mathcal{F})$ is called the Rademacher complexity of $\mathcal{F}$. Conditioning on the data $X_1, \cdots, X_n$, $\mathsf{Rad}_n(\mathcal{F})$ can be viewed as an empirical process of the random signs. Because f and random signs are bounded, we have a bounded empirical process, and hence a sub-Gaussian process.

Now let's think about what if $\mathcal{F}$ is actually a finite set with cardinality $|\mathcal{F}|$? Under this simplified setting, we can apply the following important Maximal Inequality theorem:

Theorem 5 (Maximal Inequality for sub-Gaussian random variables)**.** *Suppose* $Y_i \overset{ind}{\sim}$ *sub-Gaussian*$(0, \sigma_i^2)$ *for* $i = 1, \cdots, m$*, i.e.* $\mathbb{E}e^{\lambda Y_i} \le e^{\lambda^2\sigma_i^2/2}$*, for* $i = 1, \cdots, m$ *and* $\lambda > 0$*. If* $\sigma_1, \cdots, \sigma_m \le \sigma$*, then*

$$\begin{aligned}
\mathbb{E} \max_{1\le i\le m} Y_i &\le \sigma \left(2\log m\right)^{1/2}, \\
\mathbb{E} \max_{1\le i\le m} |Y_i| &\lesssim \sigma \left(\log 2m\right)^{1/2}
\end{aligned} \tag{13}$$

Proof.

$$\begin{aligned}
e^{\lambda \mathbb{E}\max_{1\le i\le m} Y_i} &\le \mathbb{E} e^{\lambda \max_{1\le i\le m} Y_i} \text{ by Jensen} \\
&\le \sum_{i=1}^m \mathbb{E} e^{\lambda Y_i} \le m e^{\frac{\lambda^2\sigma^2}{2}}
\end{aligned}$$

$$\Rightarrow \mathbb{E} \max_{1\le i\le m} Y_i \le \frac{\log m}{\lambda} + \frac{\lambda\sigma^2}{2}.$$

Then find out the minimizer over λ should be $\lambda^2 = 2\log m/\sigma^2$.

For the second maximal inequality, we have two proofs (one is a very clever proof, thanks to Kejie!). Augment the original sequence $Y_1, \cdots, Y_m$ to $Y_1, \cdots, Y_m, -Y_1, \cdots, -Y_m$. Then $\mathbb{E}\max_{1\le i\le m}|Y_i| = \mathbb{E}\max_{1\le i\le m}\max\{\pm Y_i\}$, followed by the first inequality. The second approach is more brute-force:

$$\begin{aligned}
\mathbb{E}\max_{1\le i\le m}|Y_i| &= \int_0^\infty \mathbb{P}\left(\max_{1\le i\le m}|Y_i| > t\right) \mathrm{d}t \\
&= \int_0^{t_0} \mathbb{P}\left(\max_{1\le i\le m}|Y_i| > t\right) \mathrm{d}t + \int_{t_0}^\infty \mathbb{P}\left(\max_{1\le i\le m}|Y_i| > t\right) \mathrm{d}t \\
&\le t_0 + 2m\int_{t_0}^\infty \frac{t}{t_0} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t.
\end{aligned}$$

Then optimize over t_0. □

The upper bound given in the first (13) is actually tight for Gaussians. I forgot to emphasize this in the class but this is quite important to remember: maxima of m sub-Gaussian random variables (each of which are $O_P(1)$) actually have order $\sqrt{\log m}$.

With Maximal Inequality, one can easily find a bound on $\mathsf{Rad}_n(\mathcal{F})$ after computing the sub-Gaussian index σ_f for every $f \in \mathcal{F}$. But now the question is what if $\mathcal{F}$ is an infinite set?

3.2.1 Metric entropy

The solution is quite simple and natural: Just find some finite set to approximate $\mathcal{F}$! The more finite elements we need, the more complex $\mathcal{F}$ is. In the field of geometric functional analysis (GFA), many people have worked on this problem and reached the consensus that a complexity measure called metric entropy is a very good way to quantify the complexity of $\mathcal{F}$. To define metric entropy, we need the following concepts.

Definition 6 (ϵ-cover and covering number)**.** Given a metric d, a set $\{\theta_1, \cdots, \theta_N : \theta_i \in \Theta, i = 1, \cdots, N\}$ is an ϵ-cover of Θ if for every $\theta \in \Theta$, there exists $i \in \{1, \cdots, N\}$ such that $d(\theta, \theta_i) \le \epsilon$. The ϵ-covering number of Θ is defined as:

$$N(\epsilon, \Theta, d) = \inf\{n \in \mathbb{N} : \exists \text{ an } \epsilon\text{-cover } \{\theta_1, \cdots, \theta_n\} \text{ of } \Theta\}. \tag{14}$$

A closely related concept is the packing number

Definition 7 (ϵ-packing and packing number)**.** Given a metric d, a set $\{\theta_1, \cdots, \theta_D : \theta_i \in \Theta, i = 1, \cdots, D\}$ is an ϵ-packing of Θ if for every $i \ne j, i, j \in \{1, \cdots, D\}$, $d(\theta_i, \theta_j) \ge \epsilon$. The ϵ-packing number of Θ is defined as:

$$D(\epsilon, \Theta, d) = \sup\{d \in \mathbb{N} : \exists \text{ an } \epsilon\text{-packing } \{\theta_1, \cdots, \theta_d\} \text{ of } \Theta\}. \tag{15}$$

Remark 8. Given Θ, its ϵ-covering and ϵ-packing are allowed to contain members which do not belong to Θ. We will see such an example in Example 2.

In fact, they are equivalent in terms of ϵ up to a constant:

Lemma 9. *For every $\epsilon > 0$,*

$$D(2\epsilon, \Theta, d) \leq N(\epsilon, \Theta, d) \leq D(\epsilon, \Theta, d).$$

Proof. The first inequality: Find a maximal 2ϵ-packing $\{\theta_1, \cdots, \theta_D\}$ and ϵ-covering $\{\theta'_1, \cdots, \theta'_N\}$. Suppose on the contrary to the statement, $D \geq N+1$. Then there must exist $i, j \in \{1, \cdots, D\}$ and $k \in \{1, \cdots, N\}$ such that $\theta_i, \theta_j \in B(\theta_k, \epsilon)$ by definition of an ϵ-covering of Θ. Thus $d(\theta_i, \theta_j) \leq 2\epsilon$ which contradicts the premise that $\{\theta_1, \cdots, \theta_D\}$ is a 2ϵ-packing.

The second inequality: Find a maximal ϵ-packing $\{\theta_1, \cdots, \theta_D\}$. Then for any $\theta \in \Theta$, there must exist $j \in \{1, \cdots, D\}$ such that $d(\theta, \theta_j) \leq \epsilon$ because otherwise $\{\theta, \theta_1, \cdots, \theta_D\}$ is also an ϵ-packing which contradicts the maximality of $\{\theta_1, \cdots, \theta_D\}$. Thus $\{\theta_1, \cdots, \theta_D\}$ is also an ϵ-covering, which implies $N(\epsilon, \Theta, d) \leq D(\epsilon, \Theta, d)$. □

Definition 10 (Metric entropy)**.** Metric entropy of Θ under metric d is simply $\log N(\epsilon, \Theta, d)$ or $\log D(\epsilon, \Theta, d)$.

Example 1. *For a bounded subset $\Theta \subseteq \mathbb{R}^d$, for every $\epsilon \in (0, 1)$*

$$\left(\frac{1}{\epsilon}\right)^d \lesssim N(\epsilon, \Theta, \|\cdot\|) \lesssim \left(\frac{1}{\epsilon}\right)^d. \tag{16}$$

Proof. The proof strategy is a classic "volume argument".

The first inequality: Find a maximal 2ϵ-packing $\{\theta_1, \cdots, \theta_D\}$. Then

$$\bigcup_{i=1}^{D} B(\theta_i, 2\epsilon) \supseteq \Theta$$

where $B(\theta, r)$ is a ball centered around θ with radius r. This is because $\{\theta_1, \cdots, \theta_D\}$ is a 2ϵ-covering of Θ. Thus

$$D\text{vol}[B(\theta_i, 2\epsilon)] \geq \text{vol}[\Theta] \Rightarrow D \geq \frac{\text{vol}[\Theta]}{\text{vol}[B(\theta_i, 2\epsilon)]} \gtrsim \left(\frac{1}{\epsilon}\right)^d.$$

The second inequality: Find a maximal ϵ-packing $\{\theta_1, \cdots, \theta_D\}$. Then

$$\bigcup_{i=1}^{D} B(\theta_i, \epsilon/2) \subseteq \widetilde{\Theta} := \{\theta \in \mathbb{R}^d : \|\theta - \Theta\| \leq \epsilon/2\}.$$

Thus

$$D\text{vol}[B(\theta_i, \epsilon/2)] \leq \text{vol}[\widetilde{\Theta}] \Rightarrow D \leq \frac{\text{vol}[\widetilde{\Theta}]}{\text{vol}[B(\theta_i, \epsilon/2)]} \lesssim \left(\frac{1}{\epsilon}\right)^d.$$

□

The above result makes intuitive sense: it reflects the curse-of-dimensionality (CoD). The complexity quantified by covering number grows exponentially with dimension d. But can we find a smaller "d" to be placed in the exponent? This will be partially answered in next section.

Example 2 (1-Lipschitz functions)**.** *Consider* $\mathcal{F} = \{f : [0,1] \to [0,1];\ f \text{ is 1-Lipschitz}\}$. *Then*

$$\log N(\epsilon, \mathcal{F}, \|\|_\infty) \lesssim \frac{1}{\epsilon}.$$

Proof. Divide $[0,1]$ into ϵ-grids. So there are $N = 1/\epsilon$ of them (I am not careful about whether the number of grids is integer or not). Denote the end points of these grids as $a_0 = 0$, $a_N = 1$ and $a_k = k\epsilon$. Define the following approximation of f as

$$\widetilde{f}(x) = \sum_{k=1}^{N} \epsilon \lfloor \frac{f(a_k)}{\epsilon} \rfloor \mathbb{1}\{x \in (a_{k-1}, a_k]\}.$$

You can easily check $\|\widetilde{f} - f\|_\infty < \epsilon$. Now we need to count how many $\widetilde{f}$ are there? In the first grid $[a_0, a_1]$ there are $1/\epsilon$ difference choices: $0, \epsilon, 2\epsilon, \cdots, 1$. However after the first grid, we have to consider an important property of $\widetilde{f}$: it is a good approximation of Lipschitz functions. So let's look at how much freedom do we have at grid k after fixing $\widetilde{f}(a_{k-1})$: By Triangle inequality and Lipschitz-ness of f

$$|\widetilde{f}(a_k) - \widetilde{f}(a_{k-1})| \le |\widetilde{f}(a_k) - f(a_k)| + |\widetilde{f}(a_{k-1}) - f(a_{k-1})| + |f(a_k) - f(a_{k-1})| \le 3\epsilon.$$

So $\widetilde{f}(a_k)$ can take at most 7 different values. Thus the total number of $\widetilde{f}$ we can have is upper bounded by

$$\frac{1}{\epsilon} 7^{1/\epsilon - 1}.$$

Thus

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim \frac{1}{\epsilon}.$$

□

3.2.2 Definition of Glivenko-Cantelli class and a sufficient condition

Glivenko-Cantelli class is the class such that uniform law of large numbers hold. In [**?**], Jon Wellner and Aad van der Vaart define GC class in the sense of strong law of large numbers, which could be quite technical. You can look at their Chapter 2 if interested. To avoid extraneous technicality, we only consider uniform WLLN.

Definition 11 (Glivenko-Cantelli (GC) class)**.** A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $\mathbb{P}|f| < \infty\ \forall f \in \mathcal{F}$ is said to belong to the strong or weak Glivenko-Cantelli (GC) class if

$$\sup_{f \in \mathcal{F}} [\mathbb{P}_n - \mathbb{P}](f) \to 0 \text{ a.s. or in } \mathbb{P}\text{-probability}$$

What conditions are sufficient for $\mathcal{F}$ to be a GC class?

Recall that we have reached the step to control $\mathbb{E}_X \mathbb{E}_\varepsilon \mathsf{Rad}_n(\mathcal{F})$ the expected Rademacher complexity of $\mathcal{F}$. Now we are going to study how to control $\mathbb{E}_X \mathbb{E}_\varepsilon \mathsf{Rad}_n(\mathcal{F})$ via maximal inequality in Theorem 5 and metric entropy. First, find a minimal ϵ-covering $\mathcal{G}_\epsilon$ of $\mathcal{F}$. Then conditioning on the data $X_1, \cdots, X_n$[2],

$$\mathbb{E}_\varepsilon \mathsf{Rad}_n(\mathcal{F}) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \le \mathbb{E}_\varepsilon \max_{g \in \mathcal{G}_\epsilon} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(X_i) \right| + \epsilon.$$

[2]Here you can compare the calculation below to what we have for Dudley's entropy integral bound in Theorem 37.

Here we are choosing norm $\|f-g\| = \frac{1}{n}\sum_{i=1}^n |f(X_i)-g(X_i)|$.

ϵ can be made arbitrarily small. For f bounded so g is also bounded between $[-B, B]$, it is not hard to see that

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i) \sim \text{sub-Gaussian}(0, \sigma_g^2)$$

where $\sigma_g^2 = \frac{1}{n^2}\sum_{i=1}^n g(X_i)^2 \lesssim \frac{1}{n}$. Thus

$$\mathbb{E}_\varepsilon \max_{g\in\mathcal{G}_\epsilon} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right| \lesssim \sqrt{\frac{\log|\mathcal{G}_\epsilon|}{n}} = \sqrt{\frac{\log N(\epsilon, \mathcal{F}, \|\cdot\|)}{n}}.$$

Thus as long as $\frac{\log N(\epsilon,\mathcal{F},\|\cdot\|)}{n} \to 0$, $\mathcal{F}$ is GC.

3.2.3 Calculating metric entropy using combinatoric dimensions

The most common combinatoric dimension of a set is via the so-called shattering number, based on which VC dimension of a set can also be defined. Shattering number is initially defined only for indicator/binary/Boolean functions.

Definition 12 (Shattering number of set of indicator functions $\mathcal{F}$)**.** Define $F_n = \{x_1, \cdots, x_n\}$ and $\Delta(\mathcal{F}, F_n) \coloneqq \{\{f(x_1), \cdots, f(x_n)\} : f \in \mathcal{F}\}$. Some people call this VC-index or projection of F_n onto $\mathcal{F}$. Shattering number is the maximal cardinality of VC index over all possible F_n:

$$s(\mathcal{F}, n) \coloneqq \sup_{F_n} \Delta(\mathcal{F}, F_n).$$

Given F_n, $\mathcal{F}$ is said to shatter F_n if $|\Delta(\mathcal{F}, F_n)| = 2^n$. Then we can define VC dimension as

$$\mathsf{VC}(\mathcal{F}) = \sup\{n \in \mathbb{N} : s(\mathcal{F}, n) = 2^n\}$$

i.e. the largest sample size such that we can find data F_n shattered by $\mathcal{F}$.

Example 3. *The class of indicator functions $\mathcal{F} = \{\mathbb{1}\{x \le c\}, c \in \mathbb{R}\}$. We have $\mathsf{VC}(\mathcal{F}) = 1$ because a set $F_2 = \{x_0, x_1\}$ without loss of generality $x_0 < x_1$ cannot be shattered by $\mathcal{F}$. We can only use $\mathcal{F}$ to pick out the subset $\{(0,0),(0,1),(1,1)\}$ which excludes $\{1,0\}$. But $\mathcal{F}$ shatters $F_1 = \{x_0\}$.*

Now let us look at a more complicated example: half spaces $\mathcal{H}_2$ in $\mathbb{R}^2$. This will be the motivating example for us to use shattering number to control the metric entropy of $\mathcal{H}_2$. We will also introduce a very useful but difficult-to-master technique called "probabilistic methods for combinatorics" [**?**]. We will only encounter a simple application of this technique.

Example 4. *We will first ask what is the upper bound of $s(\mathcal{H}_2, n)$? Look at Figure 1 below. Here $F = \{x_0, x_1, x_2, x_3\}$ so $n = 4$.*

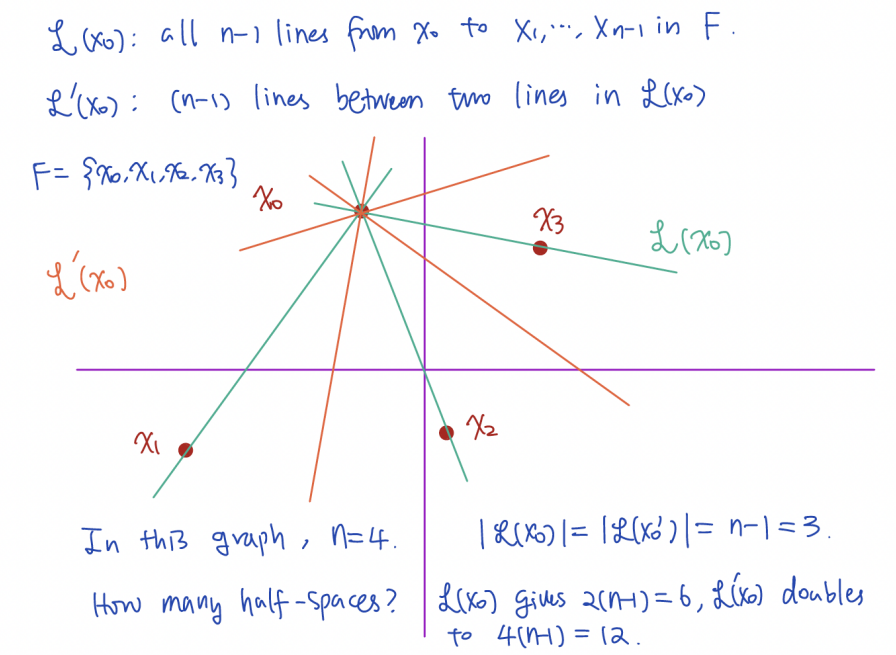

Figure 1: picture for understanding $s(\mathcal{H}_2, 4)$

We can anchor at x_0 first, and draw lines x_0-x_1, x_0-x_2, x_0-x_3, and then divide each orthant into two pieces by adding the red lines. Then there are $2 \times (4-1)$ positive half spaces (above each line) and $2 \times (4-1)$ negative half spaces (below each line). The positive and negative half spaces for each line can pick out at most two subsets from $F = \{x_0, x_1, x_2, x_3\}$ so this gives $2 \times 2 \times (4-1)$.

> *Think about why this is the largest possible subsets of F that can be picked out by the halfspaces in $\mathcal{H}_2$ going through x_0? Does cutting all the orthants one more time increase the number of subsets picked out by $\mathcal{H}_2$?*

Now generalize the above experiment from $n = 4$ to general n, giving us the upper bound $2 \times 2 \times (n-1) = 4(n-1)$. But we have n total anchor points to start with so the most conservative upper bound is

$$n4(n-1) + 1 \leq 4n^2$$

where 1 handles the possibility of $\{0, 0, 0, 0\}$. So the shattering number grows quadratically in n at least when n large, far smaller than 2^n the total number of subsets of F for general n. When $\mathsf{VC}(\mathcal{H}_2) \geq n$ though, we still have $s(\mathcal{H}_2, n) = 2^n$. But when $\mathsf{VC}(\mathcal{H}_2) < n$, a set of n data cannot be shattered any more and the shattering number grows as n^2. We will rigorously prove this phenomenon for more general cases (Sauer's lemma).

Now we will show how to derive an upper bound of the metric entropy via its shattering number upper bound n^2 via the probabilistic method for combinatorics. We first choose D half spaces $\{H_1, \cdots, H_D\}$ which is a maximal ϵ-packing of $\mathcal{H}_2$. We need an upper bound on D. Because half spaces are sets, we simply choose the following metric:

$$\|H_i - H_j\| = \mathbb{P}(H_i \Delta H_j)$$

for any probability measure $\mathbb{P}$ on $\mathbb{R}^2$. So $\mathbb{P}(H_i \Delta H_j) \geq \epsilon$ for any two members of the ϵ-packing. Now how to upper bound D? $\mathcal{H}_2$ can only pick out at most $4n^2$ subsets of any F_n with n data points. Let's ansatz all the halfspaces in the ϵ-packing pick out different subsets of $F_n = \{x_1, \cdots, x_n\}$. So $D \leq 4n^2$. But how large n can be such that the clause

> ***"all the halfspaces in the ϵ-packing pick out different subsets of $F_n = \{x_1, \cdots, x_n\}$" is satisfiable?***

That is, we need to show existence of F_n such that the above requirement is met. Here comes the ***probabilistic method****. Showing existence suffices to show the probability of the above requirement is positive. Here the probability measure is chosen such that each x_k is sampled independently for $k = 1, \cdots, n$:*

$$
\begin{aligned}
&\mathbb{P}(\textit{all the halfspaces in the } \epsilon\textit{-packing pick out different subsets of } F_n = \{x_1, \cdots, x_n\}) \\
&= 1 - \mathbb{P}(\textit{there exist two halfspaces in the } \epsilon\textit{-packing pick out the same subset from } F_n = \{x_1, \cdots, x_n\}) \\
&= 1 - \mathbb{P}(\cup_{1 \leq i < j \leq D} H_i, H_j \textit{ in the } \epsilon\textit{-packing pick out the same subset from } F_n = \{x_1, \cdots, x_n\}) \\
&\geq 1 - \binom{D}{2} \mathbb{P}(H_i, H_j \textit{ in the } \epsilon\textit{-packing pick out the same subset from } F_n = \{x_1, \cdots, x_n\}) \quad \textit{(union bound)} \\
&= 1 - \binom{D}{2} \mathbb{P}(\textit{no } x_k \in H_i \Delta H_j, \textit{ for } k = 1, \cdots, n) \\
&= 1 - \binom{D}{2} (1 - \mathbb{P}(H_i \Delta H_j))^n \\
&\geq 1 - \binom{D}{2} (1 - \epsilon)^n \\
&\geq 1 - D^2 e^{-n\epsilon} > 0 \\
&\Rightarrow e^{n\epsilon} > D^2 \Rightarrow n > \frac{2 \log D}{\epsilon}.
\end{aligned}
$$

That is, if we pick $n = \frac{2 \log D}{\epsilon}$, then it is possible to satisfy $D \leq 4n^2 = \frac{(4 \log D)^2}{\epsilon^2}$ i.e. $\frac{D}{\log^2 D} \leq \left(\frac{4}{\epsilon}\right)^2$, which further implies

$$
D \lesssim \left\{ \frac{1}{\epsilon} \log \left(\frac{1}{\epsilon} \right) \right\}^2.
$$

So with $s(\mathcal{H}_2, n) \lesssim n^2$, we can show $D(\epsilon, \mathcal{H}_2, \|\cdot\|) \lesssim \left\{ \frac{1}{\epsilon} \log \left(\frac{1}{\epsilon} \right) \right\}^2$. We will also see the more general case.

Finally, one can also ask the following question: what is the VC dimension of $\mathcal{H}_2$? In fact, $\mathsf{VC}(\mathcal{H}_2) = 3$. It has a much more general extension:

Lemma 13. *For the space of all d-dimensional half spaces $\mathcal{H}_d$ in $\mathbb{R}^d$, we have*

$$
\mathsf{VC}(\mathcal{H}_d) = d + 1.
$$

Proof. The proof is based on the following cute observation:

Fact 1. *$F_n = \{x_1, \cdots, x_n\}$ is shattered by half spaces $\{\mathbb{1}\{\beta_0 + \beta_1^\top \cdot \geq 0\}, \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^d\} \Leftrightarrow$ the columns of the design matrix*

$$
\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}_{(d+1) \times n}
$$

are linearly independent.

The intuition of the above fact is plain and simple: Since we want $\mathcal{F}$ shatters F_n, we need for every subset σ of $\{0,1\}^n$, the following system of linear equations is solvable:

$$\begin{pmatrix} a + b^\top x_1 \\ \vdots \\ a + b^\top x_n \end{pmatrix} = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{pmatrix},$$

which essentially says any $\sigma \in \{0,1\}^n$ must lie in the column space of $\boldsymbol{X}$, which is guaranteed if $d+1 \geq n$. □

Shattering number growing polynomially with n is a very general phenomenon, when the VC dimension is bounded. This is the famous Sauer's lemma.

Sauer's lemma has been proved by many different mathematicians. Sauer himself credited this problem to Paul Erdös.

Theorem 14 (Sauer's lemma)**.** *Denote d as the VC dimension of $\mathcal{F}$. Then*

$$s(\mathcal{F}, n) \leq \sum_{i=1}^{d} \binom{n}{i} \leq \left(\frac{ne}{d}\right)^d.$$

Proof. In most books, Sauer's lemma is proved by induction, which is not very interesting. I give a more constructive proof, originated from "extremal combinatorics".

To start with, let us ponder the RHS of Sauer's lemma and see what it actually entails. It is actually the cardinality of a class of subsets Δ', in which every subset member has size at most d, and every subset's subset is also in Δ'. So we ask, for VC index $\Delta \equiv \Delta(\mathcal{F}, G_n)$, can we find Δ' such that the following hold

1. $|\Delta'| = |\Delta|$;
2. If $A \in \Delta'$, then every subset of A is also in Δ';
3. The cardinality of any member $A \in \Delta'$ must be upper bounded by d.

Note that every member A of Δ or Δ' is a vector in $\{0,1\}^n$, but for every A, it also corresponds to a subset in G_n (e.g. $A = (1,1,0,\cdots,0)$ then there is a $A = \{x_0, x_1\} \subset G_n$). Because of this correspondence, we slightly abuse notation by exchangeably calling the $\{0,1\}^n$ vector and the corresponding subset in G_n both as A.

So the only task left is to construct such a Δ' and the following algorithm can do so:

- For $i = 1, \cdots, n$:
  - For $A \in \Delta$:
    - If $A \setminus \{x_i\} \notin \Delta$:
      $A \leftarrow A \setminus \{x_i\}$.

Based on the definition of this algorithm, it is obvious the requirement 1 and 2 are met because: (1) The algorithm never completely removes any member in Δ; (2) The algorithm replaces A by a smaller subset of A is that smaller subset is not in Δ.

We are left to show that the cardinality of any member of Δ' must be upper bounded by d. Note that since for any $A \in \Delta'$, any subset of A is also in Δ', there must exist Boolean function class $\mathcal{F}'$ such that $\Delta' = \Delta(\mathcal{F}', G_n)$ and $\mathcal{F}'$ shatters any $A \in \Delta'$. (Argue why this is true in your homework). If we need to show $|A| \leq d$ for $A \in \Delta'$, it suffices to show A is also shattered by $\mathcal{F}$ because $\text{VC}(\mathcal{F}) = d$. So we are left to show:

For any A shattered by $\mathcal{F}'$, A is shattered by $\mathcal{F}$.

For 3, consider we are at iteration $i \in [n]$[3]. With abuse of notation, before the "if" statement, the set is denoted as Δ; after "if", the set is denoted as Δ'.

A subset $A \subseteq G_n$ is shattered by $\mathcal{F}'$ implies that for any subset $A' \subseteq A$, we can always find $B' \in \Delta'$ such that $B' \cap A = A'$ (argue this in homework).

Then we need to show A is also shattered by $\mathcal{F}$, or equivalently there exists $B \in \Delta$ such that $B \cap A = A'$ (same argument).

Say $x_i \in A \subset \Delta'$ otherwise A will not be affected by the "if" clause during this iteration. Then divide your discussion into two parts:

(1) $x_i \in A'$, $A' \subset A$, here you can show that for $B' \in \Delta'$ with $B' \cap A = A'$, we can equate $B = B'$ and $B \in \Delta$ (argue why this is true based on the algorithm in your homework), so $\mathcal{F}$ shatters A;

(2) $x_i \notin A'$, $A' \subset A$, here you can show that there always exist $B'' \in \Delta'$ such that $B'' \cap A = A' \cup \{x_i\}$ (argue why this is true based on the algorithm in your homework). This implies $B'' \setminus \{x_i\} \in \Delta$ (argue why this is true based on the algorithm in your homework) and thus we again find $B = B'' \setminus \{x_i\}$ such that $B \cap A = A'$. □

Remark 15 (Extremal combinatorics)**.** Extremal combinatorics is a field about computing some min/max statistics in combinatorics and graph theory. The following are some typical examples of extremal problems:

- Minimum size of the largest independence set of a (triangle-free) graph G: See a recent breakthrough [**?**] by a group of undergraduate students from MIT;
- Ramsey theory: For a complete graph with edges colored with red or blue, how many vertices the graph must have to ensure the existence of a blue or red clique? Ramsey theory is also related to hardness results in Theoretical Computer Science. For example, people have conjectured that constructing a three-colored graph with d vertices with clique sizes bounded by $\log^2 d$ is computationally hard. It has been shown [**?**] that if this problem were not hard, then it would have been possible to explicitly construct a matrix satisfying the Restricted Isometry Property for $n = \log^2 d$ and the sparsity $s = \sqrt{d}$, an important condition to ensure exact recovery in sparse high-dimensional linear regression.
- Sunflower conjecture (due to Erdös and Rado): How large a family of k-sets has to be such that there exists a subfamily $\{A_1, \cdots, A_r\}$ of size r, of which the intersection between any two k-sets is the same, i.e. $A_i \cap A_j$ is independent of i, j? Erdös and Rado conjectured this

[3] In most papers, $[n]$ denotes $\{1, \cdots, n\}$ for convenience.

number should be c^k for some large constant $C > 0$. In 2019, [**?**] improved the lower bound to $(\log k)^k$ from k^k.

- Sensitivity conjecture: Every $2^{n-1}+1$-vertex subgraph of n-dimensional Boolean hypercubes (vertices adjacent iff they differ in only one coordinate) has maximum degree $\geq \sqrt{n}$; See Hao Huang's amazing two-page proof [**?**] of this decades-long open problem;
- Kadison-Singer conjecture: See Daniel Spielman and colleagues' proof [**? ? ? ?**] using spectral graph theory (eigenvalue properties of the graph Laplacian).

Apart from algebraic geometry, this is a field that witnesses a lot of new breakthroughs in recent years. It is also deeply connected with statistics and theoretical computer science. For good resources, you can check out Jacob Fox and Yufei Zhao's webpages.

Sauer's lemma has very important implication: It essentially says any set with finite VC dimension is "learnable" in the sense that its Rademacher complexity is $o(1)$. It follows from the theorem below.

Theorem 16. *For any set $\mathcal{F}$ with bounded VC dimension d, we have*

$$D(\epsilon, \mathcal{F}, ||\cdot||) \lesssim \left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)^d. \tag{17}$$

Proof. The proof is essentially the same as the proof for $\mathcal{H}_2$. Pick an ϵ-packing $\{f_1, \cdots, f_D\}$ with the same metric $\mathbb{P}(f_i \Delta f_j) \geq \epsilon$. Similarly, we can find G_n with $n \asymp \frac{\log D}{\epsilon}$. By Sauer's lemma, we immediately have

$$\begin{aligned} & D \lesssim \left(\frac{n}{d}\right)^d \lesssim \left(\frac{\log D}{\epsilon d}\right)^d \\ \Rightarrow\ & D^{1/d} \lesssim \frac{\log D}{\epsilon d} = \log D^{1/d}\frac{1}{\epsilon} \\ \Rightarrow\ & \frac{D^{1/d}}{\log D^{1/d}} \lesssim \frac{1}{\epsilon} \Rightarrow D \lesssim \left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)^d. \end{aligned}$$

□

Example 5. *The most classical example from the Glivenko-Cantelli class is the indicator functions. In fact, because indicator functions are so simple, they also belong to Donsker class, which we will cover shortly after. In this example, we will actually derive a non-asymptotic high probability bound version of the uniform law of large numbers. You will get a chance to learn another very useful exponential tail inequality (McDiarmid inequality) based on Hoeffding inequality.*

Theorem 17. *$\mathcal{F} \coloneqq \{\mathbb{1}\{\cdot \leq c\}, c \in \mathbb{R}\}$. Then we have*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}[\mathbb{P}_n - \mathbb{P}]f \leq 2\mathbb{E}Rad_n(\mathcal{F}) + \sqrt{\frac{2}{n}\log\frac{2}{\delta}}\right) \geq 1-\delta. \tag{18}$$

Proof. The proof can be divided into the following steps. The strategy is very common in writing papers.

- Denote $g(X_1, \cdots, X_n) \coloneqq \sup_{f\in\mathcal{F}} |[\mathbb{P}_n - \mathbb{P}]f|$ and decompose $g = (g - \mathbb{E}g) + \mathbb{E}g$. We first look at how well Z_n concentrates around its expectation $\mathbb{E}Z_n$, i.e. bounding

$$\mathbb{P}\left(|Z_n - \mathbb{E}Z_n| > t\right) \leq ?$$

Here because f is Boolean, $\sup_{f\in\mathcal{F}} |[\mathbb{P}_n - \mathbb{P}]f| = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\right|$ satisfies a special property called "bounded difference": i.e. for any $j = 1, \cdots, n$, we draw independent copies X_i' of X_i and compare

$$\Delta_i g = \left|\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| - \sup_{f\in\mathcal{F}} \left|\left(\frac{1}{n}\sum_{i\neq j}^n (f(X_i) - \mathbb{E}f(X)) + \frac{1}{n}(f(X_j') - \mathbb{E}f(X))\right)\right|\right| \leq \frac{1}{n}.$$

Then we can apply the McDiarmid inequality (bounded difference inequality)

Lemma 18 (McDiarmid inequality)**.** *For any function $g(x_1, \cdots, x_n)$ satisfying $|\Delta_i g| \leq c_i$ we have*

$$\mathbb{P}(|g - \mathbb{E}g| \geq t) \leq 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\} \tag{19}$$

Proof. We can decompose $g - \mathbb{E}g$ into sum of martingale differences:

$$g - \mathbb{E}g = \sum_{i=1}^n Z_i$$

where $Z_i = \mathbb{E}[g|X_1, \cdots, X_i] - \mathbb{E}[g|X_1, \cdots, X_{i-1}]$ so $\mathbb{E}[Z_i] = 0$. It is also straightforward to check that $l_i \leq Z_i \leq u_i$ for some lower and upper bounds l_i, u_i such that $u_i - l_i \leq c_i$.

Now

$$\begin{aligned}
\mathbb{P}\left(g - \mathbb{E}g > t\right) &= \mathbb{P}\left(\sum_{i=1}^n Z_i > t\right) \\
&= \mathbb{P}\left(e^{\lambda\sum_{i=1}^n Z_i} > e^{\lambda t}\right) \\
&\leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^n e^{\lambda Z_i}\right] \\
&\leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda Z_i}\mathbb{E}\left[e^{\lambda Z_n}|X_1, \cdots, X_{n-1}\right]\right] \\
\text{by Hoeffding inequality} &\leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda Z_i} e^{\frac{\lambda^2 c_n^2}{8}}\right] \\
&\leq \cdots \\
&\leq e^{-\lambda t} e^{\frac{\lambda^2}{8}\sum_{i=1}^n c_i^2}.
\end{aligned}$$

Then as before, we minimize the upper bound over λ. □

Applying (19) with c_i replaced by $1/n$, we immediately have

$$\mathbb{P}\left(|g - \mathbb{E}g| \geq t\right) \leq 2e^{-2nt^2}.$$

Eventually, by setting $\delta = 2e^{-2nt^2}$, we have $t = \sqrt{\frac{2}{n}\log\frac{2}{\delta}}$, so

$$\mathbb{P}\left(g - \mathbb{E}g \leq \sqrt{\frac{2}{n}\log\frac{2}{\delta}}\right) \geq \mathbb{P}\left(|g - \mathbb{E}g| \leq \sqrt{\frac{2}{n}\log\frac{2}{\delta}}\right) \geq 1 - \delta$$

and $\mathbb{E}g \leq 2\mathbb{E}\mathrm{Rad}_n(\mathcal{F})$. □

3.2.4 VC dimension and shattering number for non-Boolean functions?

There are now several strategies to extend the definition of shattering number and VC dimension to non-Boolean functions. The most commonly used two definitions are VC-subgraph dimension and γ-fat-shattering. [**?** **?**] (one published in Annals of Mathematics and the other published in Inventiones mathematicae, top-2 pure math journals) establish how γ-fat shattering is connected to metric entropy in $\|\cdot\|_\infty$-norm when $\mathcal{F}$ is bounded and absolutely integrable.

3.3 Bracketing number and bracketing entropy

Before we discuss rate of convergence of M-estimators, we need to introduce another type of complexity measure: bracketing numbers and bracketing entropy. These will be used frequently in later lectures.

Fix $(\mathcal{F}, ||\cdot||)$ a normed function space. $\mathcal{F} = \{f : \mathbb{X} \to \mathbb{R}\}$.

Definition 19. An ϵ-bracket of $\mathcal{F}$ is defined as

$$\left\{ \begin{array}{c} [\ell_i(\cdot), u_i(\cdot)], i = 1, \cdots, N : \\ \text{for every } f \in \mathcal{F},\ x \in \mathbb{X}, \text{ there exists } i \text{ such that } \ell_i(x) \leq f(x) \leq u_i(x) \text{ and } \|\ell - u\| \leq \epsilon \end{array} \right\}. \tag{20}$$

Then ϵ-bracketing number is

$$N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||) = \inf\{n \in \mathbb{N} : \text{ there exists an } \epsilon\text{-bracket } \{[\ell_i, u_i], i = 1, \cdots, n\} \text{ of } \mathcal{F}\}. \tag{21}$$

The bracketing entropy is simply $\log N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||)$.

We have the following relation between $N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||)$ and $N(\epsilon, \mathcal{F}, ||\cdot||)$:

Lemma 20.

$$N(\epsilon, \mathcal{F}, ||\cdot||) \leq N_{[\,]}(2\epsilon, \mathcal{F}, ||\cdot||).$$

Proof. It is obvious because 2ϵ-bracket of $\mathcal{F}$ is an ϵ-cover of $\mathcal{F}$. □

The converse is not true. But it is true in certain sense if we consider a smaller class of $\mathcal{F}$ that is related to M-estimation.

Consider $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ with (Θ, d) a metric space, satisfying the following Lipschitz-like restriction: for any $x \in \mathbb{X}$, for some envelope function F of $\mathcal{F}$.

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq d(\theta_1, \theta_2)F(x).$$

Definition 21 (Envelope)**.** An envelope F of $\mathcal{F}$ satisfies:

$$\forall\ f \in \mathcal{F}, \forall\ x \in \mathbb{X}, |f(x)| \leq F(x).$$

The minimal envelope $F^*(x) = \sup_{f\in\mathcal{F}} |f(x)|$.
Then we have

Lemma 22. *For $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ with (Θ, d) a metric space, satisfying the following Lipschitz-like restriction: for any $x \in \mathbb{X}$, for some envelope function F of $\mathcal{F}$.*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq d(\theta_1, \theta_2)F(x).$$

We have

$$N_{[\ ]}(2\epsilon||F||, \mathcal{F}, ||\cdot||) \leq N(\epsilon, \Theta, d).$$

Later in this course, we will constantly use the following quantity:

Definition 23 (Dudley's entropy integral)**.**

$$J_{[\ ]}(\delta, \mathcal{F}, ||\cdot||) \coloneqq \int_0^\delta \sqrt{\log N_{[\ ]}(\epsilon, \mathcal{F}, ||\cdot||)}\mathrm{d}\epsilon$$

3.4 Rates of convergence and asymptotic distributions of M-estimators; Donsker class; chaining and generic chaining

Consistency is usually only a first step in deriving statistical properties. In most problems, people are not satisfied with consistency. More refined characterization of an M-estimator $\widehat{\theta}_n$ includes its convergence rate to the estimand θ_0 and the asymptotic distribution of $\widehat{\theta}_n$.

We first sketch the intuition on how to derive the limiting distribution of a Z-estimator from the estimating equation perspective

$$\begin{aligned}&\widehat{\Psi}_n(\widehat{\theta}_n) \coloneqq \frac{1}{n}\sum_{i=1}^n \psi\left(X_i; \widehat{\theta}_n\right) = 0, \text{ where the target estimand } \theta_0\\ &\text{is the solution to the population expectation } \Psi(\theta) \coloneqq \mathbb{E}\left[\psi\left(X; \theta_0\right)\right] = 0.\end{aligned} \tag{22}$$

Roadmap for analyzing Z-estimators (22). We actually relax the estimating equation condition to the following:

$$\sqrt{n}\widehat{\Psi}_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1) \text{ and } \sqrt{n}\Psi(\theta_0) = 0.$$

If the estimating equations are very nonlinear, so no analytical solution is available and numerical methods need to be used, this relaxation can incorporate the case where the iterative algorithms may stop at finite number of iterations.

Then we immediately have the following:

$$\begin{aligned}o_{\mathbb{P}}(1) &= \sqrt{n}\widehat{\Psi}_n(\widehat{\theta}_n) - \sqrt{n}\Psi(\theta_0)\\ \text{centering: } &= \sqrt{n}\left(\widehat{\Psi}_n(\widehat{\theta}_n) - \Psi(\widehat{\theta}_n)\right) - \sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + \sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + \sqrt{n}\left(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)\right).\end{aligned}$$

Note that $\sqrt{n}\left(\widehat{\Psi}_n(\widehat{\theta}_n) - \Psi(\widehat{\theta}_n)\right) - \sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right)$ is an empirical process term, which we assume to be small:

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\psi(\widehat{\theta}_n) - \psi(\theta_0)] = o_{\mathbb{P}}(1). \tag{23}$$

Then we need to assume that we can perform "functional Taylor expansion" of $\Psi(\theta)$ around $\Psi(\theta_0)$: in particular, let us assume the following Fréchet differentiability assumption at the truth θ_0

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}(\theta_0)(\theta - \theta_0)\| = o\left(\|\theta - \theta_0\|\right). \tag{24}$$

$$\begin{aligned}
o_{\mathbb{P}}(1) &= o_{\mathbb{P}}(1) + \sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + \sqrt{n}\left(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)\right) \\
&\Rightarrow \sqrt{n}\left(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)\right) = -\sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + o_{\mathbb{P}}(1) \\
&\Rightarrow \sqrt{n}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0) + \sqrt{n}o_{\mathbb{P}}\left(\|\widehat{\theta}_n - \theta_0\|\right) = -\sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + o_{\mathbb{P}}(1).
\end{aligned}$$

Now we assume $\dot{\Psi}(\theta_0)$ has bounded spectrum. Then

$$\begin{aligned}
&\sqrt{n}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0) + \sqrt{n}o_{\mathbb{P}}\left(\|\widehat{\theta}_n - \theta_0\|\right) = -\sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + o_{\mathbb{P}}(1) \\
\Rightarrow\ & \sqrt{n}\|\widehat{\theta}_n - \theta_0\| = \sqrt{n}\|\dot{\Psi}(\theta_0)^{-1}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0)\| \\
&\leq \|\dot{\Psi}(\theta_0)\|^{-1}\|\sqrt{n}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0)\| \\
&= \|\dot{\Psi}(\theta_0)\|^{-1}\left\|\sqrt{n}o_{\mathbb{P}}\left(\|\widehat{\theta}_n - \theta_0\|\right) + \sqrt{n}\left(\widehat{\Psi}_n(\theta_0) - \Psi(\theta_0)\right) + o_{\mathbb{P}}(1)\right\| \\
\Rightarrow\ & \sqrt{n}\|\widehat{\theta}_n - \theta_0\| \lesssim o_{\mathbb{P}}\left(\sqrt{n}\|\widehat{\theta}_n - \theta_0\|\right) + O_{\mathbb{P}}(1) \\
\Rightarrow\ & \|\widehat{\theta}_n - \theta_0\| = O_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

With this rate of convergence, we immediately have

$$\begin{aligned}
&\sqrt{n}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0) + o_{\mathbb{P}}(1) = -\sqrt{n}(\mathbb{P}_n - \mathbb{P})\psi(\theta_0) + o_{\mathbb{P}}(1) \\
\Rightarrow\ & \sqrt{n}\dot{\Psi}(\theta_0)(\widehat{\theta}_n - \theta_0) = -\sqrt{n}(\mathbb{P}_n - \mathbb{P})\psi(\theta_0) + o_{\mathbb{P}}(1) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i;\theta_0) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \mathbb{P}\psi(\theta_0)\psi(\theta_0)^\top) \\
\Rightarrow\ & \sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\Psi}(\theta_0)^{-1}\psi(X_i;\theta_0) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \dot{\Psi}(\theta_0)^{-1}\mathbb{P}\psi(\theta_0)\psi(\theta_0)^\top\dot{\Psi}(\theta_0)^{-1}).
\end{aligned}$$

□

The above sketch is in fact a proof of the following Z-estimation theorem.

Theorem 24 (Limiting distribution for Z-estimation)**.** *Under the following conditions:*

(i) $\widehat{\Psi}_n(\widehat{\theta}_n) = o_{\mathbb{P}}(n^{-1/2})$,

(ii) $\Psi(\theta_0) = 0$,

(iii) Empirical process condition: $\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\psi(\widehat{\theta}_n) - \psi(\theta_0)] = o_{\mathbb{P}}(1)$,

(iv) $\Psi(\theta)$ *is Fréchet differentiable at* θ_0,

(v) $\dot{\Psi}(\theta_0)$ *has bounded spectrum,*

we have (1) $\|\widehat{\theta}_n - \theta_0\| = O_{\mathbb{P}}(n^{-1/2})$ *and*

$$(2)\ \sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{\Psi}(\theta_0)^{-1}\psi(X_i;\theta_0) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \dot{\Psi}(\theta_0)^{-1}\mathbb{P}\psi(\theta_0)\psi(\theta_0)^{\top}\dot{\Psi}(\theta_0)^{-1}). \quad (25)$$

Remark 25. Fréchet differentiability of $\Psi(\theta) = \mathbb{E}[\psi(X;\theta)]$ at $\theta = \theta_0$ in general can be easily checked under suitable assumptions on $\psi(x;\theta)$ and the law of $\mathbb{P}_X$, for standard parametric estimation problems, i.e. $\Theta \subset \mathbb{R}^d$ with d fixed.

We can also do the same sketch for M-estimators using optimizations.

$$\begin{aligned} &\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \widehat{M}_n(\theta) \coloneqq \sum_{i=1}^{n} m(X_i;\theta) \\ &\text{where the target estimand } \theta_0 \in \arg\max_{\theta\in\Theta} M(\theta) \coloneqq n\mathbb{E}[m(X;\theta)] \end{aligned} \quad (26)$$

The rate and distributional analysis of M-estimators are relatively more complicated.

Roadmap for analyzing M-estimators (26). Again, our goal is to show $r_n^{-1}(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0,\Sigma)$ for some appropriate r_n[4]. In finite-dimensional M-estimation problems, we can safely conjecture $r_n = n^{-1/2}$. The main idea is the following: define $t_n = r_n^{-1}(\widehat{\theta}_n - \theta_0)$ and $t_0 \sim N(0,\Sigma)$. The goal is to show $t_n \rightsquigarrow t_0$. t_n is the argmax of the following maximization problem:

$$\begin{aligned} t_n &= \arg\max_{t\in T} \widehat{M}_n(\theta_0 + \frac{1}{\sqrt{n}}t) \\ &\equiv \arg\max_{t\in T} \widehat{M}_n(\theta_0 + \frac{1}{\sqrt{n}}t) - \widehat{M}_n(\theta_0) \end{aligned}$$

say $T = \mathbb{R}^d$ or some compact subset of $\mathbb{R}^d$. Now define $U_n(t) = \widehat{M}_n(\theta_0 + \frac{1}{\sqrt{n}}t) - \widehat{M}_n(\theta_0)$, which is a stochastic process indexed by $T \subset \mathbb{R}^d$. Then

$$t_n = \arg\max_{t\in T} U_n(t).$$

Intuition: If $\{U_n(t), t \in T\}$ converges weakly to some stochastic process $\{U(t), t \in T\}$ for which $t_0 = \arg\max_{t\in T} U(t)$ (weak convergence between stochastic processes to be defined later), then we can expect $t_n \rightsquigarrow t_0$.

Definition 26 (Weak convergence between (bounded) stochastic processes)**.** For simplicity, we only deal with bounded stochastic processes, which are in general true for M-estimation problems. For the stochastic processes $\{U_n(t), t \in T\}$ and $\{U(t), t \in T\}$: $U_n, U : \Omega \to \mathcal{F}$, are both mappings from the sample space Ω of X to $\mathcal{F} \coloneqq \{f : T \to \mathbb{R}, \|f\|_\infty < \infty\}$, the space of bounded functions on T, often denoted as $\ell^\infty(T)$. Then the stochastic process $\{U_n(t), t \in T\}$ is said to converge weakly to another stochastic process $\{U(t), t \in T\}$, if for every continuous function (including linear functionals) $g : \ell^\infty(T) \to \mathbb{R}$, $g(U_n(t)) \rightsquigarrow g(U(t))$. An example of g that will be used later is $\sup_{t\in T}[\cdot] : \ell^\infty(T) \to \mathbb{R}$.

[4] In terms of how to find appropriate r_n, see Section 3.6.

We need to further impose some conditions on the empirical and population optimization problems:

$$U_n(t_n) \geq \sup_{t\in T} U_n(t) - o_{\mathbb{P}}(1) \tag{27}$$

$$U(t_0) > \sup_{t\notin G} U(t), \text{ for every open set } G \subset T \text{ s.t. } t_0 \in G \tag{28}$$

To show $t_n \rightsquigarrow t_0$, by portmanteau lemma, it is equivalent to show, for every closed $F \subset T$,

$$\limsup_n \mathbb{P}(t_n \in F) \leq \mathbb{P}(t_0 \in F).$$

First, by (27),

$$\begin{aligned}
&\mathbb{P}(t_n \in F) \leq \mathbb{P}(\sup_{t\in F} U_n(t) - \sup_{t\in T} U_n(t) + o_{\mathbb{P}}(1) \geq 0) \\
\Rightarrow\ &\limsup_n \mathbb{P}(t_n \in F) \leq \limsup_n \mathbb{P}(\sup_{t\in F} U_n(t) - \sup_{t\in T} U_n(t) \geq 0) \\
\overset{\text{Definition 26}}{\Rightarrow}\ &\limsup_n \mathbb{P}(t_n \in F) \leq \mathbb{P}(\sup_{t\in F} U(t) - \sup_{t\in T} U(t) \geq 0).
\end{aligned}$$

Finally, by (28), $\sup_{t\in F} U(t) \geq \sup_{t\in T} U(t)$ implies $t_0 \in F$ so $\limsup_n \mathbb{P}(t_n \in F) \leq \mathbb{P}(t_0 \in F)$ i.e. $t_n \rightsquigarrow t_0$. □

The above sketch is in fact a proof of the classical "argmax functional" theorem.

Theorem 27 (Argmax functional theorem)**.** *There exists a stochastic process* $\{U(t) : t \in T\}$. *Under the following conditions:*

(i) (27) *holds,*

(ii) (28) *holds,*

(iii) $\{U_n(t) : t \in T\}$ *converges weakly to* $\{U(t) : t \in T\}$ *in* $\ell^\infty(T)$,

then $t_n \rightsquigarrow t_0$.

Remark 28. A stronger argmax functional theorem can be proved by relaxing the assumptions of Theorem 27 to the following:

(i) (27) holds,

(ii) $t \mapsto -U(t)$ is lower semicontinuous (l.s.c.)[5] and t_0 is the unique maxima of $U(t)$,

(iii) $\{U_n(t) : t \in T\}$ converges weakly to $\{U(t) : t \in T\}$ in $\ell^\infty(K)$ for every compact subset $K \subset T$,

(iv) for every $\epsilon > 0$, there exists a compact set $K_\epsilon \subset T$ such that $\limsup_n \mathbb{P}(t_n \notin K_\epsilon) \leq \epsilon$ and $\mathbb{P}(t_0 \notin K_\epsilon) \leq \epsilon$.

You can read Theorem 5.56 of [**?**] or Section 3.2 of [**?**]. Lower semicontinuity is a modern condition in (non-convex) optimizations. The new condition (iv) is the so-called "tightness" condition, i.e. empirical and population optimizers are in a compact set with high probability.

[5] Equivalently, $t \mapsto U(t)$ is upper semicontinuous (u.s.c.), but l.s.c. is much more common in recent optimization literature.

3.5 Corollary on MLE

We have seen the general M-estimation theory. This theory also implies the following corollary on MLE.

Theorem 29. *For data $X_1, \cdots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$. Denote the log-likelihood function as $\ell(x;\theta)$, the score function as $S(x;\theta) = \frac{\partial \ell(x;\theta)}{\partial \theta} \equiv \dot{\ell}(x;\theta)$, and the Fisher information as $I(\theta) = \mathbb{E}_\theta[\dot{\ell}(X;\theta)^{\otimes 2}]$. Then under the set of regularity conditions similar to those in Theorem 27 or Theorem 24, we have*

$$\sqrt{n}\left(\widehat{\theta}_{MLE,n} - \theta_0\right) \rightsquigarrow N(0, I(\theta_0)^{-1}). \tag{29}$$

Remark 30. We make the following remarks on MLE:

1. $I(\theta_0)^{-1}$ is the Cramér-Rao Lower Bound. This raises the question if MLE is optimal in some sense.

2. What if the model is wrong, say the data X actually comes from a distribution $\mathbb{Q}$ instead of the posited model $\mathbb{P}_\theta$? What is MLE trying to estimate? We need to understand its estimand better:

$$\begin{aligned}
\theta_0 &= \arg\max_{\theta\in\Theta} \mathbb{E}_{X\sim\mathbb{Q}}[\log \mathrm{d}\mathbb{P}_\theta(X)] \\
&= \arg\max_{\theta\in\Theta} \int_x \log \mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{Q}(x) \\
&= \arg\max_{\theta\in\Theta} \int_x \{\log \mathrm{d}\mathbb{P}_\theta(x) - \log \mathrm{d}\mathbb{Q}(x)\}\mathrm{d}\mathbb{Q}(x) + \int_x \log \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(x) \\
&= \arg\max_{\theta\in\Theta} \int_x \log \frac{\mathrm{d}\mathbb{P}_\theta(x)}{\mathrm{d}\mathbb{Q}(x)}\mathrm{d}\mathbb{Q}(x) \\
&= \arg\min_{\theta\in\Theta} \int_x \log \frac{\mathrm{d}\mathbb{Q}(x)}{\mathrm{d}\mathbb{P}_\theta(x)}\mathrm{d}\mathbb{Q}(x) \\
&\equiv \arg\min_{\theta\in\Theta} D_{KL}(\mathbb{Q}||\mathbb{P}_\theta).
\end{aligned}$$

   Thus when the model is wrong, the MLE is trying to estimate the parameter value with which the posited model is the KL-projection of the true model $\mathbb{Q}$ onto the space of the posited model $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

3. Check Hodges' phenomenon in Section 5 for super-efficiency.

4. Finally, I did not specify when $\mathbb{E}[\dot{\ell}(X,\theta_0)\dot{\ell}(X,\theta_0)^\top] \equiv -\mathbb{E}[\ddot{\ell}(X,\theta_0)]$, which of course relies on the interchangeability between integral and derivative. But a modern treatment, mostly attributed to Lucien Le Cam, is to *only* use the following two conditions that are sufficient for a completely rigorous proof of the asymptotic normality of MLE: $\theta_0 \in \Theta \subset \mathbb{R}^d$,

   (i) *Differentiable in quadratic mean:* Denote the data generating probability measure as $\mathbb{P}_\theta$. There exists a function $\dot{\ell}(\theta_0) \equiv \dot{\ell}(X,\theta_0)$ such that as $\theta \to \theta_0$,

$$\int \left\{ \sqrt{\mathrm{d}\mathbb{P}_\theta} - \sqrt{\mathrm{d}\mathbb{P}_{\theta_0}} - \frac{1}{2}(\theta-\theta_0)^\top \dot{\ell}(\theta_0)\sqrt{\mathrm{d}\mathbb{P}_{\theta_0}} \right\}^2 = o\left(\|\theta-\theta_0\|^2\right) \tag{30}$$

(ii) *Lipschitz-type continuity:* There exists a function $\dot{\ell}$ such that, for every θ_1, θ_2 in a nbhd of θ_0

$$|\ell(x, \theta_1) - \ell(x, \theta_2)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|. \tag{31}$$

This second condition implies the objective function being Donsker.

3.6 Convergence rates via optimization

This section is for self-study. Recall that we only give the sketch for deriving distributional limits after we figured out the correct convergence rate so that $r_n^{-1}(\widehat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$. We are still left with the question how to derive convergence rates for general finite-dimensional M-estimation problem from the perspective of optimization.

Here we give the following quite general theorem, which also relies on empirical process conditions.

Theorem 31 (Rate theorem)**.** *Suppose the following:* d *a metric on* Θ*.*

(i) $\sup_{\mathrm{d}(\theta,\theta_0)\leq\delta} M(\theta) - M(\theta_0) \lesssim -n\delta^{\alpha}$ *for some* $\alpha > 0$*,*

(ii) $\mathbb{E}\left[\sup_{\mathrm{d}(\theta,\theta_0)\leq\delta} |\sqrt{n}\,(\mathbb{P}_n - \mathbb{P})\,[m(X,\theta) - m(X,\theta_0)]|\right] \lesssim \delta^{\beta}$ *for some* $\beta > 0$*.*

Then $r_n = n^{-\frac{1}{2(\alpha-\beta)}}$ *s.t.* $r_n^{-1}\mathrm{d}\left(\widehat{\theta}_n, \theta_0\right) = O_{\mathbb{P}}(1)$*.*

Proof. Without loss of generality, we choose some C such that $\log_2 C$ is an integer.

$$\begin{aligned}
&\mathbb{P}\left(\left|r_n^{-1}\mathrm{d}(\widehat{\theta}_n, \theta_0)\right| > C\right) \\
\leq &\sum_{j \geq \log_2 C} \mathbb{P}\left(2^{j-1} r_n \leq \left|\mathrm{d}(\widehat{\theta}_n, \theta_0)\right| \leq 2^j r_n\right).
\end{aligned}$$

But $\widehat{\theta}_n \in S_j \coloneqq \left\{\theta : 2^{j-1} r_n \leq |\mathrm{d}(\theta, \theta_0)| \leq 2^j r_n\right\}$ implies $\sup_{\theta \in S_j} \widehat{M}_n(\theta) - \widehat{M}_n(\theta_0) \gtrsim -o_{\mathbb{P}}(n \cdot n^{-\rho_n})$[6] for some rate ρ_n (this can be ensured by running the optimization algorithm to a certain precision). We can choose appropriate ρ_n later. Thus

$$\begin{aligned}
&\mathbb{P}\left(\left|r_n^{-1}\mathrm{d}(\widehat{\theta}_n, \theta_0)\right| > C\right) \\
\leq &\sum_{j \geq \log_2 C} \mathbb{P}\left(2^{j-1} r_n \leq \left|\mathrm{d}(\widehat{\theta}_n, \theta_0)\right| \leq 2^j r_n\right) \\
\leq &\sum_{j \geq \log_2 C} \mathbb{P}\left(\sup_{\theta \in S_j} \widehat{M}_n(\theta) - \widehat{M}_n(\theta_0) \gtrsim -o_{\mathbb{P}}(n \cdot n^{-\rho_n})\right) \\
= &\sum_{j \geq \log_2 C} \mathbb{P}\left(\sup_{\theta \in S_j} \{\widehat{M}_n(\theta) - M(\theta)\} - \{\widehat{M}_n(\theta_0) - M(\theta_0)\} + M(\theta) - M(\theta_0) \gtrsim -o_{\mathbb{P}}(n \cdot n^{-\rho_n})\right) \\
\leq &\sum_{j \geq \log_2 C} \mathbb{P}\left(\sup_{\theta \in S_j} \{\widehat{M}_n(\theta) - M(\theta)\} - \{\widehat{M}_n(\theta_0) - M(\theta_0)\} - 2^{j\alpha} n r_n^{\alpha} \gtrsim -o_{\mathbb{P}}(n \cdot n^{-\rho_n})\right)
\end{aligned}$$

[6] The extra n is because we set the scaling of $\widehat{M}_n$ as $\sum_{i=1}^n$.

$$
\begin{aligned}
&= \sum_{j \geq \log_2 C} \mathbb{P}\left(\sup_{\theta \in S_j} \{\widehat{M}_n(\theta) - M(\theta)\} - \{\widehat{M}_n(\theta_0) - M(\theta_0)\} \gtrsim 2^{j\alpha} n r_n^\alpha - o_{\mathbb{P}}(n \cdot n^{-\rho_n}) \right) \\
&\overset{\text{(ii)}}{\leq} \sum_{j \geq \log_2 C} \frac{\sqrt{n} 2^{j\beta} r_n^\beta}{n 2^{j\alpha} r_n^\alpha - n o_{\mathbb{P}}(n^{-\rho_n})} \\
&\leq \sum_{j \geq \log_2 C} \frac{1}{\sqrt{n}} \frac{2^{j\beta} r_n^\beta}{2^{j\alpha} r_n^\alpha - o_{\mathbb{P}}(n^{-\rho_n})}.
\end{aligned}
$$

So we can choose $r_n^{\beta-\alpha} = O(\sqrt{n}) \Rightarrow r_n = O(n^{-\frac{1}{2(\alpha-\beta)}})$ and ρ_n can be chosen appropriately to make sure it is dominated by r_n^α. □

Remark 32. The reason we divide the interval $[C, \infty)$ by geometric series $2^{j-1}, 2^j, 2^{j+1}, \cdots$ is obvious: we do not want the number of summands to contribute to the final sum of the probabilities. We will see later, for usual parametric models, $\beta = 1$ and for objective functions admitting second-order differentiability $\alpha = 2$, so the rate will be $n^{-1/2}$.

The bound on the empirical process is called "modulus of continuity". For M-estimation, this part essentially determines the rate of convergence (as the first condition on the optimization landscape is not really about statistics). This is because M-estimation, at least under the parametric regime (finite-dimensional parameter space with a lot of regularity), is essentially a linear problem [by localizing the statistical problem with first-order (resp. second-order) Taylor-expansion of the estimating equation (resp. optimization problem)]. In general, modulus of continuity determines the convergence rates for linear problems (this was documented in [**?** **?**]).

There are a variety of regularity conditions for M/Z-estimation. For more details, you can read relevant chapters of [**?**] (Chapters 5, 18, 19, 20). But you do not have to understand all the details of those different conditions. Just keep in mind two main things: empirical process type conditions and some smoothness conditions on the functional $\Psi(\theta)$ or $M(\theta)$ or the functions $\psi(x, \theta)$ and $m(x, \theta)$ (in the argument of θ) are unavoidable. For different applications, different sets of conditions might be more suitable. You will see one example in homework 3.

3.7 Donsker class

Now the elephant in the room is: when does the empirical process condition like (23) or Definition 26 holds? We need to introduce a new definition called the $\mathbb{P}$-Donsker class. The philosophy is again by using complexity measures like entropy. Here we focus on entropy integrals.

Definition 33 ($\mathbb{P}$-Donsker class)**.** A class $\mathcal{F}$ is $\mathbb{P}$-Donsker if the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly in $\ell^\infty(\mathcal{F})$ to a "tight" random process $\{\mathbb{G}_{\mathbb{P}} f : f \in \mathcal{F}\}$, where $\mathbb{G}_n[\cdot] = \sqrt{n}(\mathbb{P}_n - \mathbb{P})[\cdot]$.

Theorem 34 (Dudley's theorem)**.** *Weak convergence between stochastic processes* $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ *and* $\{\mathbb{G}_{\mathbb{P}} f : f \in \mathcal{F}\}$ *can be equivalently characterized by the following two conditions:*

(i) Weak convergence between every finite-dimensional distributions (fidi) of the stochastic processes:

$$(\mathbb{G}_n f_1, \cdots, \mathbb{G}_n f_k)^\top \rightsquigarrow (\mathbb{G}_{\mathbb{P}} f_1, \cdots, \mathbb{G}_{\mathbb{P}} f_k)^\top$$

for any finite set $f_1, \cdots, f_k \subset \mathcal{F}$*, for every* $k \in \mathbb{N}$*.*

(ii) Asymptotic equicontinuity: attach to $\mathcal{F}$ *a metric* d *such that for every* $\epsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{f,g \in \mathcal{F}, \mathrm{d}(f,g) < \delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| > \epsilon \right) = 0. \tag{32}$$

Since this should be covered at some point in the stochastic processes course, please read the proof of Theorem 18.14 of [**?**]. The second "asymptotic equicontinuity" condition ensures that the limiting process of $\mathbb{G}_n$ is bounded in probability (i.e. tight). Asymptotic equicontinuity is a probabilistic version of Lipschitzness condition, and if you inspect how we define a sub-Gaussian process in Theorem 37 carefully, you will see it is quite obvious why sub-Gaussianity implies asymptotic equicontinuity. In fact, in many stochastic processes textbooks, this condition is called "Lipschitz continuity" of the sample path.

We give the following two Donsker's theorems that might be useful depending on whether bracketing entropy or metric entropy is easier to derive for specific applications.

Theorem 35 (Donsker's theorem with uniform metric entropy integral)**.** *Define the uniform entropy integral as follows:*

$$J(\delta, \mathcal{F}, L_2) \coloneqq \int_0^\delta \sup_{\mathbb{Q}} \sqrt{\log N(\epsilon \|F\|_{L_2(\mathbb{Q})}, \mathcal{F}, L_2(\mathbb{Q}))} \mathrm{d}\epsilon$$

$\mathcal{F}$ *has envelope* F *such that* $\mathbb{P}F^2 < \infty$*. If* $J(1, \mathcal{F}, L_2) < \infty$*, then* $\mathcal{F}$ *is* $\mathbb{P}$*-Donsker.*

Proof. By Theorem 34, we only need to check the asymptotic equicontinuity part. First, define a new function class

$$\mathcal{G}_\delta \coloneqq \left\{ f - g : f, g \in \mathcal{F}, \|f - g\|_{L_2(\mathbb{P})} \leq \delta \right\}.$$

Obviously, $\mathcal{G}_\delta$ has an envelope $2F$. For any $\epsilon > 0$,

$$\begin{aligned} &\lim_{\delta \to 0} \limsup_n \mathbb{P}\left(\sup_{h \in \mathcal{G}_\delta} |\mathbb{G}_n h| \geq \epsilon \right) \\ \leq &\frac{1}{\epsilon} \lim_{\delta \to 0} \limsup_n \mathbb{E}\left[\sup_{h \in \mathcal{G}_\delta} |\mathbb{G}_n h| \right]. \end{aligned}$$

Now we can put on our empirical process theory hat and try to control $\mathbb{E}\left[\sup_{h \in \mathcal{G}_\delta} |\mathbb{G}_n h|\right]$. In particular, we have the following lemma:

Lemma 36. *$\mathcal{F}$ has an envelope F. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\right] \lesssim J(1,\mathcal{F},L_2)\|F\|_{L_2(\mathbb{P})}.$$

Finally, by inspecting the proof of Lemma 36,

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{G}_\delta}|\mathbb{G}_n h|\right] &\lesssim \int_0^{\sup_{f\in\mathcal{G}_\delta}\|f\|_{L_2(\mathbb{P}_n)}/\|2F\|_{L_2(\mathbb{P}_n)}} \sqrt{\log D(\epsilon\|2F\|_{L_2(\mathbb{P}_n)},\mathcal{G}_\delta,L_2(\mathbb{P}_n))}\|2F\|_{L_2(\mathbb{P}_n)}\mathrm{d}\epsilon \\ &\lesssim \int_0^{\sup_{f\in\mathcal{G}_\delta}\|f\|_{L_2(\mathbb{P}_n)}/\|F\|_{L_2(\mathbb{P}_n)}} \sup_{\mathbb{Q}}\sqrt{\log D(\epsilon\|2F\|_{L_2(\mathbb{Q})},\mathcal{G}_\delta,L_2(\mathbb{Q}))}\mathrm{d}\epsilon\|F\|_{L_2(\mathbb{P}_n)}.\end{aligned}$$

Since $J(1,\mathcal{F},L_2)$ is bounded, the integrand in the above display should be integrable, and the upper limit $\sup_{f\in\mathcal{G}_\delta}\|f\|_{L_2(\mathbb{P}_n)}\to 0$ as $\delta\to 0$, the final integral converges to 0 by dominated convergence theorem. □

Proof of Lemma 36. The proof requires some new techniques. But as before, we start with the usual symmetrization trick.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\right] \le 2\mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right].$$

Let us bound the Rademacher process term first:

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\ \lesssim& \int_0^{\sup_{f\in\mathcal{F}}\|f\|_{L_2(\mathbb{P}_n)}}\sqrt{\log D(\epsilon,\mathcal{F},L_2(\mathbb{P}_n))}\mathrm{d}\epsilon \\ =& \int_0^{\sup_{f\in\mathcal{F}}\|f\|_{L_2(\mathbb{P}_n)}/\|F\|_{L_2(\mathbb{P}_n)}}\sqrt{\log D(\epsilon'\|F\|_{L_2(\mathbb{P}_n)},\mathcal{F},L_2(\mathbb{P}_n))}\|F\|_{L_2(\mathbb{P}_n)}\mathrm{d}\epsilon' \\ \le& \|F\|_{L_2(\mathbb{P}_n)}\int_0^1\sqrt{\log D(\epsilon\|F\|_{L_2(\mathbb{P}_n)},\mathcal{F},L_2(\mathbb{P}_n))}\mathrm{d}\epsilon.\end{aligned}$$

where $L_2(\mathbb{P}_n)$ denotes the L_2-norm of a function with respect to the empirical distribution of the data i.e. $\|f\|_{L_2(\mathbb{P}_n)} = \left\{\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right\}^{1/2}$, the first line inequality follows from the famous Dudley's entropy integral bound for sub-Gaussian processes (see Theorem 37), the second line equality is due to a change of variable, and the third line inequality follows because F is an envelope so $\sup_{f\in\mathcal{F}}\|f\|_{L_2(\mathbb{P}_n)}/\|F\|_{L_2(\mathbb{P}_n)}\le 1$.

Finally, we marginalize over X:

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\right] \\ \le& 2\mathbb{E}_X\left[\|F\|_{L_2(\mathbb{P}_n)}\int_0^1\sqrt{\log D(\epsilon\|F\|_{L_2(\mathbb{P}_n)},\mathcal{F},L_2(\mathbb{P}_n))}\mathrm{d}\epsilon\right]\end{aligned}$$

$$\leq 2\mathbb{E}_X\left[\|F\|_{L_2(\mathbb{P}_n)}\right]\int_0^1 \sup_{\mathbb{Q}}\sqrt{\log D(\epsilon\|F\|_{L_2(\mathbb{Q})},\mathcal{F},L_2(\mathbb{Q}))}\mathrm{d}\epsilon$$

$$\equiv 2\mathbb{E}_X\left[\left(\frac{1}{n}\sum_{i=1}^n F(X_i)^2\right)^{1/2}\right]\int_0^1 \sup_{\mathbb{Q}}\sqrt{\log D(\epsilon\|F\|_{L_2(\mathbb{Q})},\mathcal{F},L_2(\mathbb{Q}))}\mathrm{d}\epsilon$$

$$\leq 2\underbrace{\left(\mathbb{E}_X F(X)^2\right)^{1/2}}_{=:\|F\|_{L_2(\mathbb{P})}}\underbrace{\int_0^1 \sup_{\mathbb{Q}}\sqrt{\log D(\epsilon\|F\|_{L_2(\mathbb{Q})},\mathcal{F},L_2(\mathbb{Q}))}\mathrm{d}\epsilon}_{=:J(1,\mathcal{F},L_2)}.$$

□

Theorem 37 (Dudley's entropy integral bound for sub-Gaussian processes; chaining argument)**.** *Consider a stochastic process $(X_t, t\in T)$ indexed by a separable[7] metric space (T,d). Further we assume it is a sub-Gaussian process, which generalizes Gaussian process by only inheriting its tail probability:*

$$\mathbb{P}\left(|X_t - X_s| > u\right) \leq 2\exp\left(-\frac{u^2}{2\mathrm{d}(s,t)^2}\right),\ \textit{for any } s,t\in T \textit{ and } u>0. \tag{33}$$

Then for some fixed $t_0\in T$,

$$\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] \lesssim \sum_{k=1}^N 2^{-k}\sqrt{\log D(\epsilon_k,T,\mathrm{d})} \lesssim \int_0^{D/2}\sqrt{\log D(\epsilon,T,\mathrm{d})}\mathrm{d}\epsilon$$

where D is the diameter of T, i.e. $D\coloneqq \sup_{t,s\in T}\mathrm{d}(s,t)$.

Proof. In the proof, we will go at a very slow pace to see how Richard Dudley's chaining argument was developed. Chaining argument is one of the center pillars of modern machine learning theory. X_{t_0} can be treated as a constant, e.g. 0.

We assume T to be countable. If T is not countable, then we can simply take a dense countable subset of T by its separability. In the calculations for Glivenko-Cantelli theorem, we take a maximal ϵ_1-packing, say T_1, so $|T_1| = D(\epsilon_1,T,\mathrm{d})$. We then essentially did the following calculations: For every t, we call $\pi_1(t)$ the projection of t onto the ϵ_1-packing T_1, then

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] &= \mathbb{E}\left[\sup_{t\in T}|X_t - X_{\pi_1(t)} + X_{\pi_1(t)} - X_{t_0}|\right]\\ &\leq \underbrace{\mathbb{E}\left[\sup_{t\in T}|X_{\pi_1(t)} - X_{t_0}|\right]}_{I_0} + \underbrace{\mathbb{E}\left[\sup_{t\in T}|X_t - X_{\pi_1(t)}|\right]}_{I_1}.\end{aligned}$$

Now we analyze each term separately. I_0 is obviously a familiar term, and can be handled by maximal inequality because there are only finitely many different $X_{\pi_1(t)} - X_{t_0}$ for $t\in T$ as $\pi_1(t)\in T_1$. I_1, though, is not as nice as it looks. In the Glivenko-Cantelli case, the corresponding X_t, due to the scaling of $\frac{1}{n}$ instead of the scaling of $\frac{1}{\sqrt{n}}$ in the Donsker case, is not a random variable asymptotically (i.e. deterministic quantity). But in the Donsker case or in the context of this theorem, X_t

[7] A space which has a dense countable subset.

is a random variable. In the above decomposition, $\sup_{t\in T} \mathrm{d}(t, \pi_1(t)) \le \epsilon$, which implies certain high-probability uniform closeness between X_t and $X_{\pi_1(t)}$ by sub-Gaussianity; otherwise T_1 is not maximal.

Now let us suppose T is even finite. We can choose a smaller $\epsilon_2 < \epsilon_1$ and a finer (higher-resolution) maximal ϵ_2-packing T_2 and further decompose the above display as

$$\begin{aligned}&\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right]\\ \le\ &\mathbb{E}\left[\sup_{t\in T}|X_{\pi_1(t)} - X_{t_0}|\right] + \mathbb{E}\left[\sup_{t\in T}|X_{\pi_2(t)} - X_{\pi_1(t)}|\right] + \mathbb{E}\left[\sup_{t\in T}|X_t - X_{\pi_2(t)}|\right].\end{aligned}$$

We repeat the above scheme N times: since T is finite, there must exist a finite N such that the maximal ϵ_N-packing $T_N \equiv T$ so $\pi_N(t) \equiv t$. Hence

$$\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] \le \sum_{k=1}^{N} \mathbb{E}\left[\sup_{t\in T}|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}|\right].$$

Now each term within the expectation is a supremum over finitely many possible choices $\pi_k(t)$ and $\pi_{k+1}(t)$ because T_k and T_{k+1} are finite. Thus we can use maximal inequality to control each summand as:

$$\begin{aligned}&\mathbb{E}\left[\sup_{t\in T}|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}|\right]\\ &\lesssim \mathrm{d}(\pi_k(t), \pi_{k-1}(t))\sqrt{\log|T_{k-1}||T_k|}\\ &\le 2\epsilon_{k-1}\sqrt{\log|T_k|^2} = 2\epsilon_{k-1}\sqrt{2\log D(\epsilon_k, T, \mathrm{d})}\end{aligned}$$

where the second line inequality follows from $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ being sub-Gaussian with variance proxy $\mathrm{d}(\pi_k(t), \pi_{k-1}(t))^2$. Now if we choose $\epsilon_{k-1} \lesssim \epsilon_k$ (but remember $\epsilon_k < \epsilon_{k-1}$), we have

$$\mathbb{E}\left[\sup_{t\in T}|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}|\right] \lesssim \epsilon_k\sqrt{\log D(\epsilon_k, T, \mathrm{d})}.$$

Thus

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] &\lesssim \sum_{k=1}^{N}\epsilon_k\sqrt{\log D(\epsilon_k, T, \mathrm{d})} \le \sum_{k=1}^{N}\int_{\epsilon_{k+1}}^{\epsilon_k}\sqrt{\log D(\epsilon, T, \mathrm{d})}\mathrm{d}\epsilon\\ &\le \int_{\epsilon_{N+1}}^{\epsilon_1}\sqrt{\log D(\epsilon, T, \mathrm{d})}\mathrm{d}\epsilon \le \int_0^{\epsilon_1}\sqrt{\log D(\epsilon, T, \mathrm{d})}\mathrm{d}\epsilon.\end{aligned}$$

How should we choose ϵ_k such that $\epsilon_{k-1} \lesssim \epsilon_k$ but $\epsilon_k < \epsilon_{k-1}$? For ϵ_0, we do not have much choice than the diameter D because T_0 is a maximal ϵ_0-packing of T containing t_0, which means $\sup_t \mathrm{d}(t, t_0) \lesssim \epsilon_0$. The fastest rate that ϵ_k converges to 0 is then exponential to make sure ϵ_k and ϵ_{k-1} are on equal order. Thus we can choose $\epsilon_k = D2^{-k}$.

Finally, remember we are still assuming T is finite. However, does it really matter? When T is countable, we can simply choose any finite subset $T_J^\dagger \subset T$ with cardinality J and establish

$$\mathbb{E}\sup_{t\in T_J^\dagger}|X_t - X_{t_0}| \lesssim \int_0^{D/2}\sqrt{\log D(\epsilon, T, \mathrm{d})}\mathrm{d}\epsilon.$$

Since the upper bound does not depend on J, we have

$$\lim_{J\to\infty} \mathbb{E}\sup_{t\in T_J^\dagger} |X_t - X_{t_0}| \lesssim \int_0^{D/2} \sqrt{\log D(\epsilon, T, \mathrm{d})}\mathrm{d}\epsilon.$$

□

In fact, building on an important result by Aad van der Vaart and Jon Wellner [**?**], in 2014, Victor Chernozhukov proved one of the most powerful versions of the maximal inequality as far as I know:

Lemma 38 (Theorem 5.2 of [**?**])**.** *Suppose there exists* $\sigma > 0$ *such that* $\sup_{f\in\mathcal{F}} \mathbb{P}f^2 \le \sigma^2 \le \mathbb{P}F^2$*, set* $\delta \coloneqq \dfrac{\sigma}{\|F\|_{L_2(\mathbb{P})}}$ *and* $C \coloneqq \sqrt{\mathbb{P}\max_{1\le i\le n} F(X_i)^2}$*. Then*

$$\mathbb{E}\left[\sup_{h\in\mathcal{G}_\delta} |\mathbb{G}_n h|\right] \lesssim J(\delta, \mathcal{F}, L_2)\|F\|_{L_2(\mathbb{P})} + \frac{CJ(\delta, \mathcal{F}, L_2)^2}{\delta^2\sqrt{n}}. \tag{34}$$

The proofs of the above lemma by both papers are quite simple and pedagogical. You should read the proofs on your own.

Theorem 39 (Donsker's theorem with bracketing entropy integral)**.** $\mathcal{F}$ *has envelope* F *such that* $\mathbb{P}F^2 < \infty$*. If* $J_{[\,]}(1, \mathcal{F}, L_2(\mathbb{P})) < \infty$*, then* $\mathcal{F}$ *is* $\mathbb{P}$*-Donsker.*

Proof. The proof idea is drastically similar so we omit the proof here. The maximal inequality part is a bit technical and you can look at Lemma 19.34 of [**?**]. □

Bracketing entropies have been derived for many important function classes (Lipschitz, Sobolev, Hölder, Besov, ...) in [**?**]. Also see many examples in [**?**].

For instance, for the class of α-smooth functions:

$$\mathcal{F} = \left\{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \lfloor\alpha\rfloor\text{-differentiable and } f^{(\lfloor\alpha\rfloor)} \text{ is } (\alpha - \lfloor\alpha\rfloor)\text{-Hölder continuous}\right\}$$

of which the bracketing entropy is know to be $(1/\epsilon)^{d/\alpha}$. Then it is easy to see that its bracketing entropy integral is finite if $\alpha > d/2$. For example, if $d = 20$, then the underlying function needs more than 10 derivatives to be Donsker.

3.8 Generic chaining; Talagrand γ_2 functional

Dudley's chaining argument, as powerful as it is, does not achieve tight result for Gaussian processes. By Sudakov's inequality:

$$\mathbb{E}\sup_{t\in T} X_t \gtrsim \sup_{\epsilon>0} \epsilon\sqrt{\log D(\epsilon, T, \mathrm{d})},$$

so there is a gap between this lower bound and the upper bound of Dudley's entropy integral theorem.

I will not have time to cover all the amazing things about Gaussian processes, including Slepian's and Sudakov-Fernique's Gaussian comparison inequalities, Gaussian interpolation techniques, and Gordon's inequality. You should learn these materials from Chapter 7 of [**?**].

Talagrand made a monumental contribution by closing this gap via an improved chaining technique, called "generic chaining". Roman Vershynin's book [**?**] explained how he interpreted the reasoning behind the generic chaining technique. But it is still not super clear to me, in particular why Talagrand chose the size of the maximal packings to be 2^{2^k}. Here is my own understanding.

First, let us recall Dudley's entropy integral bound before turned into an integral (for short, we call it Dudley's $\mathcal{D}_2$ functional):

$$\mathcal{D}_2 \coloneqq \sum_{k=1}^{N} \sup_{t\in T} \sqrt{\log D(\epsilon_k, T, \mathrm{d})} \cdot \mathrm{d}(t, \pi_k(t)) \lesssim \sum_{k=1}^{N} \sup_{t\in T} \sqrt{\log D(\epsilon_k, T, \mathrm{d})} \cdot 2^{-k}.$$

So Dudley fix the resolution $\epsilon_k \asymp 2^{-k}$ and then obtain the maximal ϵ_k-packing T_k for every resolution from $k = 0, \cdots, N$.

Talagrand, however, took a dual viewpoint: he fixed the size of some maximal packing T_k^a to be 2^{2^k} then looked for the corresponding resolution level ϵ_k. Of course, the larger the size of T_k^a, the higher the resolution we are trying to approximate T. So the "pre-dual" of $\mathcal{D}_2$ is

$$\mathcal{D}_2^\dagger \coloneqq \sum_{k=1}^{N} \sup_{t\in T} \sqrt{\log D(\epsilon_k, T, \mathrm{d})} \cdot \mathrm{d}(t, \pi_k(t)) \le \sum_{k=1}^{N} \sup_{t\in T} \sqrt{\log 2^{2^k}} \cdot \mathrm{d}(t, \pi_k(t)).$$

Talagrand's first move is to pull sup in front of the summation, obtaining:

$$\mathcal{D}_2^* \coloneqq \sup_{t\in T} \sum_{k=1}^{N} \sqrt{\log D(\epsilon_k, T, \mathrm{d})} \cdot \mathrm{d}(t, \pi_k(t)) \le \sup_{t\in T} \sum_{k=1}^{N} \sqrt{\log 2^{2^k}} \cdot \mathrm{d}(t, \pi_k(t)) \lesssim \sup_{t\in T} \sum_{k=1}^{N} 2^{k/2} \mathrm{d}(t, \pi_k(t)).$$

Obviously, $\mathcal{D}_2^* \le \mathcal{D}_2^\dagger$. So if one can show $\mathcal{D}_2^*$ is an upper bound, it must be tighter than $\mathcal{D}_2^\dagger$.

Since the above quantity is important historically, people have given it a name: Talagrand's γ_2 functional

$$\gamma_2 \coloneqq \inf_{(T_k^a)} \sup_{t\in T} \sum_{k=1}^{N} 2^{k/2} \mathrm{d}(t, \pi_k(t)) \tag{35}$$

where Talagrand even optimized over all possible maximal packings satisfying the 2^{2^k} size condition.

Theorem 40 (Talagrand's generic chaining theorem)**.** *For a zero-mean sub-Gaussian process $(X_t, t \in T)$ with (T, d), we have*

$$\mathbb{E}\left[\sup_{t\in T} |X_t|\right] \lesssim \gamma_2. \tag{36}$$

Proof. As before, with the notation $X_{\pi_0}(t) \equiv 0$,

$$X_t = \sum_{k=1}^{N} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}.$$

Since now $\sup_{t\in T}$ is outside the summation $\sum_{k=1}^N$, we need to instead show a uniform bound on

$$|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}|$$

uniformly over t. By sub-Gaussianity, the variance proxy for $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ is upper bounded by $\mathrm{d}(\pi_k(t), \pi_{k-1}(t))$ up to constant. We want to bound

$$\mathbb{P}\left(\sup_{t\in T}\sum_{k=1}^N |X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > u\right) \lesssim ?$$

Observation: the event $\left\{\sup_{t\in T}\sum_{k=1}^N |X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > uR\right\}$, with $R = \sup_{t\in T}\sum_{k=1}^N \sqrt{\log|T_k^a|}\mathrm{d}(t,\pi_k(t))$, implies

$$\left\{|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > u\sqrt{\log|T_k^a|}\mathrm{d}(t,\pi_k(t)),\ \text{for some } t\in T \text{ and for some } k = 1,\cdots,N\right\}.$$

This is because the contra-positive of the above statement obviously holds. Hence, by union bound,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in T}\sum_{k=1}^N |X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > uR\right)\\
&\le \mathbb{P}\left(|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > u\sqrt{\log|T_k^a|}\mathrm{d}(t,\pi_k(t)),\ \text{for some } t\in T \text{ and for some } k = 1,\cdots,N\right)\\
&\le \sum_{t\in T, 1\le k\le N} \mathbb{P}\left(|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > u\sqrt{\log|T_k^a|}\mathrm{d}(t,\pi_k(t))\right)\\
&\le \sum_{k=1}^N |T_k^a||T_{k-1}^a| \exp\left\{-\frac{u^2\log|T_k^a|\mathrm{d}(t,\pi_k(t))^2}{2\mathrm{d}(t,\pi_k(t))^2}\right\} = \sum_{k=1}^N |T_k^a||T_{k-1}^a| \exp\left\{-\frac{u^2\log|T_k^a|}{2}\right\}.
\end{aligned}$$

To make sure the summation is addable, we need to choose the size T_k^a such that

$$|T_k^a|^2 \lesssim e^{\frac{u^2\log|T_k^a|}{2}}$$

which gives us $|T_k^a| = 2^{2^k}$.

Finally (you should check the following calculations on your own): with some constant $c > 0$ chosen appropriately,

$$\begin{aligned}
\mathbb{E}\left[\frac{\sup_{t\in T}|X_t|}{R}\right] &= \int_0^\infty \mathbb{P}\left(\sup_{t\in T}|X_t| > uR\right)\mathrm{d}u\\
&= c + \int_c^\infty \mathbb{P}\left(\sup_{t\in T}|X_t| > uR\right)\mathrm{d}u\\
&\lesssim c + \int_c^\infty \sum_{k=1}^N 2^{3\cdot 2^{k-1}} \exp\left\{-u^2 2^{k-1}\right\}\mathrm{d}u\\
&= c + \int_c^\infty \sum_{k=1}^N 2^{3\cdot 2^{k-1}} \exp\left\{-u^2 2^{k-1}\right\}\mathrm{d}u
\end{aligned}$$

$$
\begin{aligned}
&= c + \sum_{k=1}^{N} 2^{3\cdot 2^{k-1}} \int_c^\infty e^{-u^2 2^{k-1}} \mathrm{d}u \\
&= c + \sum_{k=1}^{N} 2^{3\cdot 2^{k-1}} 2^{-\frac{k}{2}} \underbrace{\int_{c2^{k/2}}^\infty e^{-\frac{u^2}{2}} \mathrm{d}u}_{\text{Tail of } N(0,1)} \\
&\lesssim c + \sum_{k=1}^{N} 2^{-\frac{k}{2}} 2^{3\cdot 2^{k-1}} e^{-\frac{c^2 2^k}{2}}.
\end{aligned}
$$

So we can choose c such that $2^{3\cdot 2^{k-1}} e^{-\frac{c^2 2^k}{2}} = O(1)$ and the proof is done. □

Talagrand invented more than one technique to improve Dudley's chaining argument. An alternative method to generic chaining is "majorizing measure" [**?**], which is the tool to show that for Gaussian process $(X_t, t \in T)$,

$$\mathbb{E} \sup_{t\in T} X_t \asymp \gamma_2.$$

3.9 Some concrete examples

Concrete examples of M/Z-estimation can be found in Chapter 5 of [**?**], including sample median, robust estimation, and etc. We will leave this part for self-study.

But we will consider the following somewhat more abstract application of how to apply what we have learnt so far.

Theorem 41. *Consider M-estimation with objective function $m(x,\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, Θ open. Moreover, almost surely at the true law $\mathbb{P}$, $\theta \mapsto m(x,\theta)$ is differentiable at $\theta_0 \in \Theta$ with derivative $\dot{m}(x,\theta_0)$, and that for $F \in L_2(\mathbb{P})$, we have*

$$|m(x,\theta_1) - m(x,\theta_2)| \le F(x)\|\theta_1 - \theta_2\|$$

for all θ_1, θ_2 in $\mathcal{N}(\theta_0)$. Suppose

$$\mathbb{P}m(\theta) - \mathbb{P}m(\theta_0) = \frac{1}{2}(\theta-\theta_0)^\top V(\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$$

where V is symmetric and negative definite. If $\widehat{\theta}_n \to_{\mathbb{P}} \theta_0$, then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n(\dot{m}(\theta_0)) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, V^{-1}\mathbb{P}(\dot{m}(\theta_0)\dot{m}(\theta_0)^\top)V^{-1}).$$

Proof. Consider the optimization perspective. Define

$$U_n(t) = \sum_{i=1}^{n} m(X_i, \theta_0 + tn^{-1/2}) - m(X_i,\theta_0).$$

We need to find its limiting stochastic process $U(t)$. Guess: $U(t) = t^\top \mathbb{G}_n(\dot{m}(\theta_0)) + \frac{1}{2}t^\top V t$. □

4 Concentration of measures

4.1 Talagrand inequalities

We start off with the following Bernstein inequality without proof. This theorem has been proved many times in different books; e.g. see Chapter 3.1 of [**?**].

Theorem 42 (Bernstein inequality)**.**

1. *X a mean-zero random variable with MGF satisfying for some $\nu > 0$*

$$\mathbb{E}e^{\lambda X} \leq \exp\left(\nu(e^{\lambda} - 1 - \lambda)\right), \lambda > 0.$$

Then for any $u \geq 0$, define $h(x) = (1+x)\log(1+x) - x$, we have

$$\begin{gathered}\mathbb{P}\left(X \geq u\right) \leq \exp\left(-\nu h(u/\nu)\right) \leq \exp\left(-\frac{3u}{4}\log\left(1 + \frac{2u}{3\nu}\right)\right) \leq \exp\left(-\frac{u^2}{2\nu + 2u/3}\right), \\ \mathbb{P}\left(X \geq \sqrt{2\nu u} + u/3\right) \leq e^{-u}.\end{gathered} \tag{37}$$

2. *$X \sim$ sub-exponential$(0, \sigma, \alpha)$ i.e.*

$$\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{for } |\lambda| < \frac{1}{\alpha}.$$

Then for any $u > 0$,

$$\mathbb{P}\left(X \geq u\right) \leq \exp\left(-\frac{u^2}{2\sigma^2 + 2\alpha u}\right). \tag{38}$$

Bernstein inequality can be tighter than Hoeffding inequality even for certain bounded random variables: e.g. Bernoulli.

Corollary 1.

1. *For $X_1, \cdots, X_n$ independent random variables with $|X_i| \leq B$ almost surely for some $B > 0$ for all $i = 1, \cdots, n$. Denote $\sigma^2 \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^2$.*

$$\mathbb{E}e^{\lambda\sum_{i=1}^n X_i} \leq \exp\left(\frac{n\sigma^2}{B^2}(e^{\lambda B} - 1 - \lambda B)\right), \lambda > 0.$$

Then for any $u \geq 0$,

$$\begin{gathered}\mathbb{P}\left(\sum_{i=1}^n X_i \geq u\right) \leq \exp\left(-\frac{n\sigma^2}{B^2}h\left(\frac{uB}{n\sigma^2}\right)\right) \leq \exp\left(-\frac{3u}{4B}\log\left(1 + \frac{2uB}{3n\sigma^2}\right)\right) \leq \exp\left(-\frac{u^2}{2n\sigma^2 + 2Bu/3}\right) \\ \mathbb{P}\left(\sum_{i=1}^n X_i \geq \sqrt{2n\sigma^2 u} + Bu/3\right) \leq e^{-u}.\end{gathered} \tag{39}$$

2. *For $X_1, \cdots, X_n$ independent sub-exponential random variables sub-exponential$(0, \sigma_i, \alpha)$.*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq u\right) \leq \exp\left(-\frac{u^2}{2n\sigma^2 + 2\alpha u}\right). \tag{40}$$

Remark 43. For bounded random variables, we make the following comparison between Bernstein and Hoeffding: since Bernstein inequality uses the variance information, it is generally tighter than Hoeffding inequality, which only uses the bound.

sub-Gaussian and sub-exponential random variables are special cases of a more general class of random variables – sub-Weibull. You may find relevant results in the paper [**?**], which essentially rewrites many results from Martin Wainwright's high-dimensional statistics book and Victor Chernozhukov's maximal inequality [**?**] in a more unified framework. I did not have time to cover sub-Gaussian and sub-exponential random variables in detail: e.g. what is the so-called Orlicz/gauge norm. You may learn all these details from [**?**].

Talagrand inequality is essentially Bernstein inequality for stochastic processes $(X_t, t \in T)$. In this course, we will prove the following version of Talagrand inequality.

Theorem 44 (Talagrand inequality: Bousquet upper tail)**.** *Let $X_1, \cdots, X_n$ be independent random variables. Let $\mathcal{F}$ be a countable set of functions f such that $\|f\|_\infty \leq B$ for some $B > 0$ and $\mathbb{E}f(X_1) = \cdots = \mathbb{E}f(X_n) = 0$. Define*

$$Z_n \coloneqq \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i), \sigma^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n} \sup_{f\in\mathcal{F}} \mathbb{E}f(X_i)^2, \nu_n \coloneqq 2B\mathbb{E}Z_n + n\sigma^2.$$

Then for any $\lambda > 0$,

$$\begin{aligned} \mathbb{E}e^{\lambda(Z_n - \mathbb{E}Z_n)} &\leq e^{\nu_n(e^\lambda - 1 - \lambda)}, \\ \mathbb{P}\left(Z_n \geq \mathbb{E}Z_n + u\right) &\leq \exp\left(-\frac{u^2}{2\nu_n + 2Bu/3}\right), \\ \mathbb{P}\left(Z_n \geq \mathbb{E}Z_n + \sqrt{2\nu_n u} + Bu/3\right) &\leq e^{-u}. \end{aligned} \tag{41}$$

The conclusion also holds if we define $Z_n \coloneqq \sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} f(X_i)|$.

Before proving Theorem 44, we make some comments. Talagrand inequality is actually a series of inequalities that concern with the following philosophy/phenomenon as explained by Talagrand himself:

> *For a (nonlinear) function of n random variables $f(X_1, \cdots, X_n)$, when f is sufficiently smooth/regular/nice, $f(X_1, \cdots, X_n)$ should be roughly a constant, highly concentrating around $\mathbb{E}f(X_1, \cdots, X_n)$.*

Proof. Without loss of generality, we take $B = 1$. It is sufficient to prove

$$\log \mathbb{E}e^{\lambda(Z_n - \mathbb{E}Z_n)} \leq \nu_n\left(e^\lambda - 1 - \lambda\right).$$

The high probability bound follows from the log-MGF bound.

To prove the log-MGF bound, we need to introduce the following important quantity and the so-called "entropy method" based on log-Sobolev inequality: for any measurable function $f(X_1, \cdots, X_n) \geq 0$, the entropy functional with respect to the measure μ is

$$\mathsf{Ent}_\mu f := \mathbb{E}_\mu f \log f - \mathbb{E}_\mu f \log \mathbb{E}_\mu f.$$

We gather the following facts about $\mathsf{Ent}_\mu f$:

Lemma 45 (Facts about $\mathsf{Ent}_\mu f$)**.**

1. $\mathsf{Ent}_\mu f$ *is homogeneous of degree 1:* $\mathsf{Ent}_\mu \lambda f = \lambda \mathsf{Ent}_\mu f$*; by homogeneity, without loss of generality, we assume* $\mathbb{E}_\mu f \equiv 1$ *in the proof.*

2. *Variational characterization of entropy functional*

$$\mathsf{Ent}_\mu f \equiv \sup \left\{ \int f g \mathrm{d}\mu : \int e^g \mathrm{d}\mu \leq 1 \right\} \tag{42}$$

$$\mathsf{Ent}_\mu f \equiv \inf \left\{ \int [f \log f - (\log t + 1) f + t] \mathrm{d}\mu : t \geq 0 \right\} \tag{43}$$

   As a corollary of (43)*, one has*

$$\mathsf{Ent}_\mu e^f \equiv \inf \left\{ \int \phi(-(f - t)) e^f \mathrm{d}\mu : t \in \mathbb{R} \right\} \tag{44}$$

   where $\phi(x) = e^x - 1 - x$*.*

3. $\mathsf{Ent}_\mu f$ *tensorizes in the following sense:*

$$\mathsf{Ent}_\mu f \leq \sum_{i=1}^n \int_{x_{-i}} \mathsf{Ent}_{\mu_i} f \mathrm{d}\mu(x). \tag{45}$$

4. *If* $\|f\|_\infty \leq 1$*, then for any* $\lambda > 0$

$$\mathsf{Ent}_\mu e^{\lambda f} \leq \int \phi(-\lambda f) e^{\lambda f} \mathrm{d}\mu \leq \underbrace{\frac{\phi(-\lambda) e^\lambda}{e^\lambda - \frac{1}{2}}}_{\equiv m(\lambda)} \int f \cdot \left(e^{\lambda f} + \frac{1}{2} f - 1 \right) \mathrm{d}\mu. \tag{46}$$

Proof of Lemma 45.

1. Simple algebra.

2. We can use the following version of Young's inequality: for $x \in \mathbb{R}$ and $y \geq 0$, $xy \leq y \log y - y + e^x$ and equality holds when $y = e^x$. Hence take x as g and y as f, we have, setting $g^* = \log f$

$$\max_g \int f g \mathrm{d}\mu = \int \left(f \log f - f + e^{g^*} \right) \mathrm{d}\mu \equiv \int (f \log f - f + f) \, \mathrm{d}\mu \equiv \mathsf{Ent}_\mu f.$$

3. Since $x \mapsto x \log x$ is convex for $x \geq 0$,

$$\begin{aligned}\int f \log f \mathrm{d}\mu &= \inf_t \int \left[f \log f - t \log t - (f-t)(t \log t)'\right] \mathrm{d}\mu \\ &= \inf_t \int \left[f \log f - t \log t - (f-t)(1 + \log t)\right] \mathrm{d}\mu \\ &= \inf_t \int \left[f \log f - (\log t + 1) + t\right] \mathrm{d}\mu.\end{aligned}$$

4. For any g, define

$$g_1(x) = \log \frac{e^{g(x)}}{\int_{x_1} e^{g(x)} \mathrm{d}\mu_1}, g_i(x) = \log \frac{\int_{x_1,\cdots,x_{i-1}} e^{g(x)} \mathrm{d}\mu_1 \cdots \mathrm{d}\mu_{i-1}}{\int_{x_1,\cdots,x_i} e^{g(x)} \mathrm{d}\mu_1 \cdots \mathrm{d}\mu_i}, i \geq 2.$$

   Then

$$g \leq g - \log \int e^{g(x)} \mathrm{d}\mu = \sum_{i=1}^n g_i.$$

   Hence take the optimal g^*,

$$\mathsf{Ent}_\mu f = \int f g^* \mathrm{d}\mu \leq \sum_{i=1}^n \int f g_i^* \mathrm{d}\mu \leq \sum_{i=1}^n \int \mathsf{Ent}_{\mu_i} f \mathrm{d}\mu.$$

5. Taking $t = 0$ in (44), we have

$$\begin{aligned}\mathsf{Ent}_\mu e^{\lambda f} &\leq \int \phi(-\lambda f) e^{\lambda f} \mathrm{d}\mu \\ &= \int \left(e^{-\lambda f} - 1 + \lambda f\right) e^{\lambda f} \mathrm{d}\mu \\ &= \int \left(1 - e^{\lambda f} + \lambda f e^{\lambda f}\right) \mathrm{d}\mu.\end{aligned}$$

   Next we observe $\frac{1 - e^{\lambda f} + \lambda f e^{\lambda f}}{f\left(e^{\lambda f} + \frac{1}{2} f - 1\right)}$ attains maxima at $f \equiv 1$, at which the maxima equals $m(\lambda) \equiv \frac{\phi(-\lambda) e^\lambda}{e^\lambda - \frac{1}{2}}$.

□

Lemma 45 are about the properties of the entropy functional. For this version of Talagrand inequality, we also have quite a special f – suprema of sum of independent random variables. We further utilize this special structure. First, it is quite obvious Z_n is subadditive in the following sense

Lemma 46 (Subadditivity of Z_n)**.** *Suppose* $\|f\|_\infty \leq 1$ *and also* $\mathbb{E} f(X) = 0$*. Define*

$$Z_n^{(k)} := \max_{f \in \mathcal{F}} \sum_{i=1, i \neq k}^n f(X_i).$$

Then we have

$$Z_n - Z_n^{(k)} \le 1, (n-1)Z_n \le \sum_{k=1}^n Z_n^{(k)}, Z_n - \mathbb{E}_k Z_n \le Z_n - Z_n^{(k)} \le 1 \tag{47}$$

where $\mathbb{E}_k[\cdot] \equiv \mathbb{E}[\cdot|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n]$. *Also, there exists random variables* Y_k *such that* $Y_k \le Z_n - Z_n^{(k)}$ *and* $\mathbb{E}_k Y_k = 0$ *and*

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}_k Y_k^2 \le \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathbb{E} f(X_i)^2.$$

Proof of Lemma 46. $Z_n - Z_n^{(k)} \le 1$ is obvious from the definition. Take $f^* = \arg\max_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ and $f_k^* = \arg\max_{f\in\mathcal{F}} \sum_{i=1, i\neq k}^n f(X_i)$. Then

$$\sum_{k=1}^n Z_n - Z_n^{(k)} \le \sum_{k=1}^n \left(\sum_{i=1}^n f^*(X_i) - \sum_{i=1, i\neq k}^n f^*(X_i) \right) = \sum_{k=1}^n f^*(X_k) = Z_n$$
$$\Rightarrow (n-1)Z_n \le \sum_{k=1}^n Z_n^{(k)}.$$

Next:

$$\begin{aligned} Z_n - \mathbb{E}_k Z_n &\le Z_n - \mathbb{E}_k \left[\sum_{i=1, i\neq k}^n f_k^*(X_i) + f_k^*(X_k) \right] \\ &= Z_n - \sum_{i=1, i\neq k}^n f_k^*(X_i) - \mathbb{E}_k[f_k^*(X_k)] = Z_n - Z_n^{(k)} \le 1. \end{aligned}$$

Finally, we take $Y_k \equiv f_k^*(X_k)$. □

Combining the above two lemma, we have

$$\begin{aligned} \mathsf{Ent}_{\mathbb{P}} e^{\lambda Z_n} &\overset{\text{Lemma 45.3}}{\le} \sum_{k=1}^n \mathbb{E}\left[\mathsf{Ent}_{\mathbb{P}_k} e^{\lambda Z_n} \right] \\ &\overset{\text{Lemma 45.1}}{\le} \sum_{k=1}^n \mathbb{E}\left[e^{\lambda \mathbb{E}_k Z_n} \mathsf{Ent}_{\mathbb{P}_k} e^{\lambda(Z_n - \mathbb{E}_k Z_n)} \right] \\ &\overset{\text{Lemma 45.4, 46}}{\le} m(\lambda) \sum_{k=1}^n \mathbb{E}\left[e^{\lambda \mathbb{E}_k Z_n} \mathbb{E}_k \left((Z_n - \mathbb{E}_k Z_n) e^{\lambda(Z_n - \mathbb{E}_k Z_n)} + \frac{1}{2}(Z_n - \mathbb{E}_k Z_n)^2 - (Z_n - \mathbb{E}_k Z_n) \right) \right] \\ &= m(\lambda) \sum_{k=1}^n \mathbb{E}\left[e^{\lambda \mathbb{E}_k Z_n} \left(\mathbb{E}_k (Z_n - \mathbb{E}_k Z_n) e^{\lambda(Z_n - \mathbb{E}_k Z_n)} + \frac{1}{2}\mathbb{E}_k (Z_n - \mathbb{E}_k Z_n)^2 \right) \right] \\ &= m(\lambda) \sum_{k=1}^n \mathbb{E}\left[\mathbb{E}_k (Z_n - \mathbb{E}_k Z_n) e^{\lambda Z_n} + \frac{1}{2}\mathbb{E}_k (Z_n - \mathbb{E}_k Z_n)^2 e^{\lambda \mathbb{E}_k Z_n} \right] \end{aligned}$$

$$
\begin{aligned}
&\overset{\text{Jensen}}{\leq} m(\lambda)\sum_{k=1}^{n}\mathbb{E}\left[\mathbb{E}_k Z_n e^{\lambda Z_n} - \mathbb{E}_k Z_n \mathbb{E}_k e^{\lambda Z_n} + \frac{1}{2}\mathbb{E}_k(Z_n - \mathbb{E}_k Z_n)^2 \mathbb{E}_k e^{\lambda Z_n}\right] \\
&= m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}(n-1) Z_n e^{\lambda Z_n} - \mathbb{E}\sum_{k=1}^{n}\left[\mathbb{E}_k Z_n \mathbb{E}_k e^{\lambda Z_n} - \frac{1}{2}\mathbb{E}_k(Z_n - \mathbb{E}_k Z_n)^2 \mathbb{E}_k e^{\lambda Z_n}\right]\right) \\
&\overset{\text{Lemma 46}}{\leq} m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\sum_{k=1}^{n} Z_n^{(k)} \mathbb{E}_k e^{\lambda Z_n} - \mathbb{E}\sum_{k=1}^{n}\left[\mathbb{E}_k Z_n \mathbb{E}_k e^{\lambda Z_n} - \frac{1}{2}\mathbb{E}_k(Z_n - \mathbb{E}_k Z_n)^2 \mathbb{E}_k e^{\lambda Z_n}\right]\right) \\
&= m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\left[\sum_{k=1}^{n}\left(Z_n^{(k)} - \mathbb{E}_k Z_n + \frac{1}{2}\mathbb{E}_k(Z_n - \mathbb{E}_k Z_n)^2\right)\mathbb{E}_k e^{\lambda Z_n}\right]\right) \\
&= m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\left[\sum_{k=1}^{n}\left(Z_n^{(k)} - \mathbb{E}_k Z_n + \frac{1}{2}\mathbb{E}_k(Z_n - \mathbb{E}_k Z_n)^2\right) e^{\lambda Z_n}\right]\right) \\
&\overset{\text{Lemma 46}}{\leq} m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\left[\sum_{k=1}^{n}\left(-\mathbb{E}_k(Z_n - Z_n^{(k)}) + \frac{1}{2}\mathbb{E}_k(Z_n - Z_n^{(k)})^2\right) e^{\lambda Z_n}\right]\right) \\
&\overset{\text{Lemma 46}}{\leq} m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\left[\sum_{k=1}^{n}\left(-\underbrace{\mathbb{E}_k Y_k}_{\equiv 0} + \frac{1}{2}\mathbb{E}_k Y_k^2\right) e^{\lambda Z_n}\right]\right) \\
&\equiv m(\lambda)\left(\mathbb{E} Z_n e^{\lambda Z_n} + \mathbb{E}\left[\frac{1}{2}\sum_{k=1}^{n}\mathbb{E}_k Y_k^2 e^{\lambda Z_n}\right]\right).
\end{aligned}
$$

Consequently, we have, by homogeneity,

$$
\begin{aligned}
\mathsf{Ent}_{\mathbb{P}} e^{\lambda(Z_n - \mathbb{E}Z_n)} &\leq m(\lambda)\left(\mathbb{E} Z_n e^{\lambda(Z_n - \mathbb{E}Z_n)} + \mathbb{E}\left[\frac{1}{2}\sum_{k=1}^{n}\mathbb{E}_k Y_k^2 e^{\lambda(Z_n - \mathbb{E}Z_n)}\right]\right) \\
&= m(\lambda)\left(\mathbb{E}(Z_n - \mathbb{E}Z_n) e^{\lambda(Z_n - \mathbb{E}Z_n)} + \mathbb{E}\left[\left(\frac{1}{2}\sum_{k=1}^{n}\mathbb{E}_k Y_k^2 + \mathbb{E}Z_n\right) e^{\lambda(Z_n - \mathbb{E}Z_n)}\right]\right) \\
&= m(\lambda)\left(\mathbb{E}(Z_n - \mathbb{E}Z_n) e^{\lambda(Z_n - \mathbb{E}Z_n)} + \left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\mathbb{E}\left[e^{\lambda(Z_n - \mathbb{E}Z_n)}\right]\right).
\end{aligned}
$$

We make a note here: the second line in the above display is the so-called modified log-Sobolev inequality, where log-Sobolev inequality is for Gaussian random variables. The above calculation, which we called entropy method, is also called Herbst argument.

Finally, some calculus trick. Define $L(\lambda) \coloneqq \log \mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)}$ as the log-MGF. Then

$$
L'(\lambda) = \left\{\mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)}\right\}^{-1} \mathbb{E}(Z_n - \mathbb{E}Z_n) e^{\lambda(Z_n - \mathbb{E}Z_n)}
$$

We observe that for the above inequality,

$$
\begin{aligned}
\text{LHS} &= \mathbb{E}\lambda(Z_n - \mathbb{E}Z_n) e^{\lambda(Z_n - \mathbb{E}Z_n)} - \mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)} \log \mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)} \\
&= \lambda \mathbb{E}(Z_n - \mathbb{E}Z_n) e^{\lambda(Z_n - \mathbb{E}Z_n)} - L(\lambda)\mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)} \\
&= (\lambda L'(\lambda) - L(\lambda))\mathbb{E} e^{\lambda(Z_n - \mathbb{E}Z_n)}
\end{aligned}
$$

and

$$\text{RHS} = m(\lambda)\left(L'(\lambda)\mathbb{E}e^{\lambda(Z_n-\mathbb{E}Z_n)} + \left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\mathbb{E}e^{\lambda(Z_n-\mathbb{E}Z_n)}\right).$$

Hence

$$\begin{aligned}
&(\lambda - m(\lambda))L'(\lambda) - L(\lambda) \le m(\lambda)\left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\\
\Rightarrow\ &\frac{\lambda e^\lambda - \frac{1}{2}\lambda - 1 + e^\lambda - \lambda e^\lambda}{e^\lambda - \frac{1}{2}}L'(\lambda) - L(\lambda) \le \frac{(1 - e^\lambda + \lambda e^\lambda)}{e^\lambda - \frac{1}{2}}\left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\\
\Rightarrow\ &\frac{e^\lambda - \frac{1}{2}\lambda - 1}{e^\lambda - \frac{1}{2}}L'(\lambda) - \frac{e^\lambda - \frac{1}{2}}{e^\lambda - \frac{1}{2}}L(\lambda) \le \frac{(1 - e^\lambda + \lambda e^\lambda)}{e^\lambda - \frac{1}{2}}\left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\\
\Rightarrow\ &\left(e^\lambda - \frac{1}{2}\lambda - 1\right)L'(\lambda) - \left(e^\lambda - \frac{1}{2}\right)L(\lambda) \le \left(1 - e^\lambda + \lambda e^\lambda\right)\left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\\
\Rightarrow\ &\left(\frac{L(\lambda)}{e^\lambda - \frac{1}{2}\lambda - 1}\right)' \le \left(\frac{1}{2}n\sigma^2 + \mathbb{E}Z_n\right)\left(-\frac{\lambda}{e^\lambda - \frac{1}{2}\lambda - 1}\right)'\\
\Rightarrow\ &L(\lambda) \le (e^\lambda - 1 - \lambda)\nu_n.
\end{aligned}$$

□

Apart from the above Bousquet upper tail, we also have the following:

Theorem 47 (Talagrand inequality: Klein/Rio lower tail)**.** *Under the same conditions as in Theorem 44, we have for any $\lambda > 0$,*

$$\mathbb{E}e^{-\lambda(Z_n-\mathbb{E}Z_n)} \le e^{\frac{\nu_n}{16}(e^{4\lambda}-1-4\lambda)}. \tag{48}$$

You may also find some other versions of Talagrand inequalities. For example:

Theorem 48. *Again let $f \equiv f(X_1, \cdots, X_n)$. If f satisfies*

$$f(x) - f(y) \le \sum_{i=1}^n c_i(x)\mathbb{1}\{x_i \ne y_i\}, \forall x, y$$

then

$$\begin{aligned}
\mathbb{P}\left(f - \mathbb{E}f \ge u\right) &\le \exp\left(-u^2/\|2\sum_{i=1}^n c_i^2\|_\infty\right)\\
\mathbb{P}\left(f - \mathbb{E}f \le -u\right) &\le \exp\left(-u^2/\|2\sum_{i=1}^n c_i^2\|_\infty\right).
\end{aligned}$$

You can find the proof in Chapter 4 of [**?**], which heavily relies on techniques related to Wasserstein distance, which we will not have time to cover. But the above theorem has an important corollary:

Corollary 2. *$X_1, \cdots, X_n \overset{ind.}{\sim} [0, 1]$. If f is convex, then $f(X_1, \cdots, X_n)$ is $\|\|\nabla f\|_{\ell^2}^2\|_\infty$-sub-Gaussian.*

This corollary is quite useful if you want to control norms, if they are convex.

4.1.1 Applications

My favorite application of Talagrand inequality is exponential inequality for second-order U-statistic. You should try to go through the proof (Chapter 3 of [**?**]) at least once in your life time.

Another classical application of Talagrand inequality is the following Dvoretzky-Kiefer-Wolfowitz non-asymptotic concentration of empirical measures.

Theorem 49 (Dvoretzky-Kiefer-Wolfowitz)**.**

$$\mathbb{P}\left(\|\sqrt{n}\left(\mathbb{P}_n[\mathbb{1}_.] - \mathbb{P}[\mathbb{1}_.]\right)\|_\infty > u\right) \lesssim \exp\left(-\frac{u^2}{C}\right).$$

for some constant $C > 0$.

For its proof, see homework 3.

For other applications, you will find them when you read papers in statistics and machine learning.

4.2 Anti-concentration inequality

Talagrand inequality is essentially a "concentration of measure" phenomenon [**?**]. In the probability and nonparametric statistics literature, you will hear another related concept called anti-concentration. But it is not about showing the reverse of the above concentration result. Anti-concentration inequalities are related to deriving lower bounds, which are often much more difficult than upper bounds (Sourav Chatterjee's recent paper [**?**] establishes a somewhat more general framework).

Theorem 50 (Anti-concentration of Gaussian maxima [**?**])**.** *Let $(X_t, t \in T)$ be a zero-mean unit-variance Gaussian process indexed by a metric space (T, d). Then*

$$\begin{aligned}\sup_{x\in\mathbb{R}} \mathbb{P}\left(\left|\sup_{t\in T} X_t - x\right| \le \varepsilon\right) &\lesssim \varepsilon\left(\mathbb{E}\sup_{t\in T} X_t \vee 1\right),\\ \sup_{x\in\mathbb{R}} \mathbb{P}\left(\left|\sup_{t\in T} |X_t| - x\right| \le \varepsilon\right) &\lesssim \varepsilon\left(\mathbb{E}\sup_{t\in T} |X_t| \vee 1\right).\end{aligned} \tag{49}$$

Thus anti-concentration is saying Gaussian maxima does not concentrate "near" any particular number; whereas concentration of measure is saying that Gaussian maxima is "near" its expectation. But the scales of "nearness" between "concentration" and "anti-concentration" are quite different. Anti-concentration inequality is very useful in constructing confidence sets; see the application in [**?**].

5 Hodges' phenomenon

We have seen in Theorem 29 that under regularity conditions, MLE asymptotically achieves the Cramér-Rao bound (but recall that we have a sharp lower bound for only unbiased estimators, which in general are not MLE). One may ask, is MLE optimal? If not, does there exist any estimator that beats MLE? If yes, in what sense MLE is optimal.

Consider the simple problem of estimating μ by $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$. Obviously, $\bar{X}$ is the MLE of μ. Joseph Hodges asked if it is possible to obtain a better estimator than $\bar{X}$. He proposed the following estimator:

$$\widehat{\mu}_{\text{super}} = \begin{cases} \bar{X} & \bar{X} > n^{-1/4}, \\ 0 & \bar{X} < n^{-1/4}. \end{cases} \tag{50}$$

The "merit" of $\widehat{\mu}_{\text{super}}$ was argued based on the following analysis: Denote $Z \sim N(0,1)$ and ϕ and Φ to be its PDF and CDF.

- At $\theta = 0$,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}_{\theta=0}\left(|\bar{X}| < n^{-1/4}\right) \\ &= \lim_{n\to\infty} \mathbb{P}_{\theta=0}\left(|\sqrt{n}(\bar{X} - 0)| < n^{1/4}\right) \\ &= \lim_{n\to\infty} \mathbb{P}(|Z| < n^{1/4}) \\ &= \lim_{n\to\infty} \Phi(n^{1/4}) - \Phi(-n^{1/4}) \\ &= \Phi(\infty) - \Phi(-\infty) = 1. \end{aligned}$$

  Hence $\widehat{\mu}_{\text{super}}$ is asymptotically unbiased for $\theta = 0$ with an asymptotic variance equal to 0.

- At any $\theta \neq 0$,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(|\bar{X}| \geq n^{-1/4}\right) = \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}(\bar{X} - \theta) \geq n^{1/4} - n^{1/2}\theta\right) + \mathbb{P}_\theta\left(\sqrt{n}(\bar{X} - \theta) \leq -n^{1/4} - n^{1/2}\theta\right)$$
$$= \{1 - \Phi(-\infty)\} + \Phi(-\infty) = 1.$$

  Conditioning on the event $\{|\bar{X}| \geq n^{-1/4}\}$, $\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta) \overset{d}{\to} N(0,1)$. Then

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta) \leq t\right) = \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta) \leq t \middle| |\bar{X}| \geq n^{-1/4}\right) \mathbb{P}_\theta\left(|\bar{X}| \geq n^{-1/4}\right)$$
$$= \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}(\bar{X} - \theta) \leq t \middle| |\bar{X}| \geq n^{-1/4}\right) \mathbb{P}_\theta\left(|\bar{X}| \geq n^{-1/4}\right)$$
$$= \mathbb{P}(Z \leq t)$$

  where we used "conditioning on an event with probability 1 is equivalent to not conditioning on any event" in the last equality.

To summarize, when $\theta \neq 0$, $\widehat{\mu}_{\text{super}}$ is asymptotically the same random variable as T_n whereas when $\theta = 0$, $\widehat{\mu}_{\text{super}}$ estimate 0 with 100% precision (no variability at all). Hence Hodges concluded that his estimator $\widehat{\mu}_{\text{super}}$ was more efficient than R. A. Fisher's MLE.

The above arguments have one flaw – all the stochastic limit results above are **pointwise**. As emphasized in [**?**], limit results are only useful if it can provide guidance on finite sample performance. **Pointwise** limit results are in general not useful under this principle. To see this, let's rephrase these stochastic limit results at $\theta \neq 0$ in "ε-δ language".

- Given any fixed $\theta \neq 0$, for every $\varepsilon > 0$, there exists a finite integer $n(\theta, \varepsilon)$ such that for every $n \geq n(\theta, \varepsilon)$, $\left|\mathbb{P}_\theta(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta) \leq t) - \mathbb{P}_{Z\sim N(0,1)}(Z \leq t)\right| \leq \varepsilon$ for all $t \in \mathbb{R}$. Why we cannot have a common $n(\varepsilon)$ independent of θ such that the above asymptotic normality result holds? For two different $\theta_1 \neq \theta_2$, for normal approximation to have the same error $\varepsilon > 0$, we need

  $$\begin{aligned}
  &\mathbb{P}_{\theta_i}(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta_i) \leq t) - \mathbb{P}(Z \leq t) \\
  &= \mathbb{P}_{\theta_i}\left(\sqrt{n}(\bar{X} - \theta_i) \leq t, |\bar{X}| > n^{-1/4}\right) + \mathbb{P}_{\theta_i}\left(-\sqrt{n}\theta_i \leq t, |\bar{X}| \leq n^{-1/4}\right) - \mathbb{P}(Z \leq t) \\
  &= \mathbb{P}\left(Z \leq t, |n^{-1/2}Z + \theta_i| > n^{-1/4}\right) - \mathbb{P}(Z \leq t) + \mathbb{P}\left(|n^{-1/2}Z + \theta_i| \leq n^{-1/4}\right) \mathbb{1}\left\{-n^{1/2}\theta_i \leq t\right\} \\
  &= -\mathbb{P}\left(Z \leq t, |Z + n^{1/2}\theta_i| \leq n^{1/4}\right) + \mathbb{P}\left(|Z + n^{1/2}\theta_i| \leq n^{1/4}\right) \mathbb{1}\left\{-n^{1/2}\theta_i \leq t\right\}
  \end{aligned}$$

  to be absolutely bounded by ε. Fix the tolerated error threshold $\epsilon = 4 \times 10^{-4}$. Take $t = 0$, $\theta_1 = -1/(10.001)$, then it becomes

  $$\begin{aligned}
  &\left|\mathbb{P}_{\theta_1}(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta_1) \leq 0) - \mathbb{P}(Z \leq 0)\right| \\
  &= \left|-\mathbb{P}\left(Z \leq 0, |Z + n^{1/2}\theta_1| \leq n^{1/4}\right) + \mathbb{P}\left(|Z + n^{1/2}\theta_1| \leq n^{1/4}\right) \mathbb{1}\left\{-n^{1/2}\theta_1 \leq 0\right\}\right| \\
  &= \left|-\mathbb{P}\left(Z \leq 0, |Z - n^{1/2}/(10.001)| \leq n^{1/4}\right) + \mathbb{P}\left(|Z - n^{1/2}/(10.001)| \leq n^{1/4}\right) \mathbb{1}\left\{n^{1/2}/(10.001) \leq 0\right\}\right| \\
  &= \mathbb{P}\left(Z \leq 0, |Z - n^{1/2}/(10.001)| \leq n^{1/4}\right) \\
  &= \mathbb{P}\left(-n^{1/4} + n^{1/2}/(10.001) \leq Z \leq 0\right),
  \end{aligned}$$

  which is non-increasing with n. Obviously, if we take $n = n(\varepsilon, \theta_1) = 10^4$, this error bound $\leq \epsilon = 4 \times 10^{-4}$, where equality holds when $n = n(\varepsilon, \theta_1)$. Now we are going to show that there exists a θ_2 such that at $n = 10^4$ and $t = 0$, normal approximation error is not even remotely close to $\varepsilon = 4 \times 10^{-4}$. Then for different **unknown** parameter θ, we need different sample sizes to make sure our normal approximation error can be tolerated. Let's simply take $\theta_2 = -1/50$. Now the approximation error becomes

  $$\begin{aligned}
  &\left|\mathbb{P}_{\theta_2}(\sqrt{n}(\widehat{\mu}_{\text{super}} - \theta_2) \leq 0) - \mathbb{P}(Z \leq 0)\right| \\
  &= \mathbb{P}\left(-n^{1/4} + n^{1/2}/(50) \leq Z \leq 0\right) \approx 0.5.
  \end{aligned}$$

  Remember θ is unknown. Thus even if we collect $n = 10^4$ samples, the error of normally approximating $\widehat{\mu}_{\text{super}}$ may still be way off depending on what the true θ is.

However, the above stochastic limit results, though we have not made explicit, are **uniform** instead of **pointwise**. By our normality assumption, without using CLT, we do not rely on any asymptotics and conclude that $\sqrt{n}(\bar{X} - \theta) \sim N(0, 1)$ uniformly over any $\theta \in \mathbb{R}$.

5.0.1 Squared risk of $\widehat{\mu}_{\text{super}}$

The above analysis can also be done if we use squared risk to compare $\widehat{\mu}_{\text{super}}$ and $\bar{X}$. Define the loss as $\ell(\widehat{\theta}_n, \theta) \coloneqq (\theta - \widehat{\theta}_n)^2$. Then $R_\theta(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2$. We want to compare $R_\theta(S_n)$ and $R_\theta(\bar{X})$. First, we observe that $nR_\theta(\bar{X}) = n\mathsf{var}_\theta(\bar{X}) = 1$, independent of the **unknown** parameter θ. Since

$nR_\theta(\bar{X})$ is a constant, it is natural to compare the risk after scaled by n. However, due to hard thresholding, $nR_\theta(\widehat{\mu}_{\text{super}})$ will depend on θ, as we will show below.

$$
\begin{aligned}
& nR_\theta(\widehat{\mu}_{\text{super}}) = n\mathbb{E}_\theta\left(\widehat{\mu}_{\text{super}} - \theta\right)^2 = n\mathbb{E}_\theta\left(\bar{X} - \theta\right)^2 \mathbb{1}\{|\bar{X}| > n^{-1/4}\} + n\theta^2\mathbb{E}_\theta\mathbb{1}\{|\bar{X}| \le n^{-1/4}\} \\
&= n\mathbb{E}\left[\frac{Z^2}{n}\mathbb{1}\{|n^{-1/2}Z + \theta| > n^{-1/4}\}\right] + n\theta^2\mathbb{P}\{|n^{-1/2}Z + \theta| \le n^{-1/4}\} \\
&= \mathbb{E}\left[Z^2\mathbb{1}\{|Z + n^{1/2}\theta| > n^{1/4}\}\right] + n\theta^2\mathbb{P}\{|Z + n^{1/2}\theta| \le n^{1/4}\} \\
&= 1 - \mathbb{E}\left[Z^2\mathbb{1}\{|Z + n^{1/2}\theta| \le n^{1/4}\}\right] + n\theta^2\mathbb{P}\{|Z + n^{1/2}\theta| \le n^{1/4}\} \\
&= 1 + \left\{n\theta^2 - \mathbb{E}\left[Z^2\middle|\{|Z + n^{1/2}\theta| \le n^{1/4}\}\right]\right\}\mathbb{P}\{|Z + n^{1/2}\theta| \le n^{1/4}\}.
\end{aligned}
$$

Notice that $Z|\{a \le Z \le b\}$ is called a truncated normal distribution and we need to compute its second moment to finish the above calculation. In particular, we have

$$
\mathbb{E}\left[Z^2|a \le Z \le b\right] = 1 + \frac{a\phi(a) - b\phi(b)}{\Phi(b) - \Phi(a)}.
$$

Remark 51. The wikipedia page for truncated normal distribution will be a valuable resource in the future: truncated normal distribution.

Thus

$$
nR_\theta(\widehat{\mu}_{\text{super}}) = 1 + \left\{\begin{array}{c} \left(n\theta^2 - 1\right)\left[\Phi(n^{1/4} - n^{1/2}\theta) - \Phi(-n^{1/4} - n^{1/2}\theta)\right] \\ -(-n^{1/4} - n^{1/2}\theta)\phi(-n^{1/4} - n^{1/2}\theta) + (n^{1/4} - n^{1/2}\theta)\phi(n^{1/4} - n^{1/2}\theta) \end{array}\right\}.
$$

Now we can compare $nR_\theta(\widehat{\mu}_{\text{super}})$ as a function of θ over different n's in Figure 2 using the following R code:

```
risk <- function (x, n) {
    n * (n^(-1) + (x^2 - n^(-1)) * (pnorm(n^(1/4) - n^(1/2) * x) - pnorm(-
        n^(1/4) - n^(1/2) * x)) - ((- n^(1/4) - n^(1/2) * x) * dnorm(- n^(
       1/4) - n^(1/2) * x) - (n^(1/4) - n^(1/2) * x) * dnorm(n^(1/4) - n^(
       1/2) * x)) / n)
}

pdf('Risk.pdf')
curve(risk(x, n = 100), from = -1, to = 1, xlab = expression(theta), ylab
   = 'Risk', ylim = c(0, 200), n = 500, lwd = 2)
curve(risk(x, n = 1000), add = TRUE, col = 'red', n = 500, lwd = 2)
curve(risk(x, n = 10000), add = TRUE, col = 'dodgerblue2', n = 500, lwd =
   2)
curve(risk(x, n = 100000), add = TRUE, col = 'forestgreen', n = 500, lwd =
    2)
legend('topright', c('n=100', 'n=1000', 'n=10000', 'n=100000'), text.col =
    c('black', 'red', 'dodgerblue2', 'forestgreen'), col = c('black', 'red
   ', 'dodgerblue2', 'forestgreen'))
dev.off()
```

In particular, we observe that no matter how large the sample size becomes, there always exists $\theta \in \mathbb{R}$ such that $nR_\theta(\widehat{\mu}_{\text{super}})$ is higher than any pre-specified threshold. Such θ will change with sample size n.

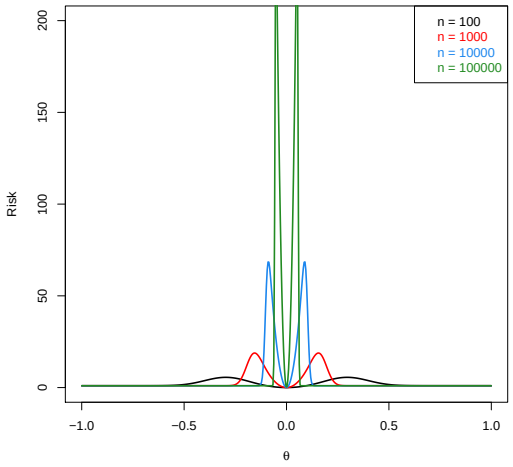

Figure 2: $nR_\theta(\widehat{\mu}_{\text{super}})$ for different sample sizes $n = 100, 1000, 10000, 100000$.

Hodges' phenomenon is also closely related to the problem of statistical inference (building valid confidence intervals) after model selection/data preprocessing/data dredging/exploratory data analysis. See some early papers by Hannes Leeb [**?** **?**]. Historically, Jianqing Fan proved that Lasso or other related high-dimensional sparsity-pursuit optimization-based linear regression techniques have a so-called "oracle variable selection" property, i.e. as sample size $n \to \infty$, under mild conditions on the covariates, these methods select the true contributing covariates with probability converging to 1. So the confidence intervals for those "estimated non-zero regression coefficients" should be valid because they are the "true non-zero regression coefficients". Then Leeb and colleagues affirmatively showed that Fan's argument for valid inference is far from being useful in practice because his results are only for point-wise consistency. Valid inference requires uniform consistency. This back-and-forth foreshadows the direction of selective inference, which is still a hot research topic today.

6 Asymptotic optimality of MLE in parametric models: Years of development by Lucien Le Cam

In this section, we cover important works that Lucien Le Cam and Jaroslav Hajék have done to show the optimality of MLE, taking the Hodges' phenomenon into account.

We have seen that showing hardness is an art for hypothesis testing problems. Asymptotic optimality of MLE in parametric models was a central topic from 1950's to 1970's, and the modern point of view was established by Lucien Le Cam. Vladimir Spokoiny [**?**] took on the challenge of rewriting Le Cam's theory in non-asymptotic terms, which are the more dominating style of modern statistics and machine learning.

6.1 Proof of Remark 30.3

Recall that for MLE to converge to a limiting normal distribution, we consider the following quadratic expansion of the log-likelihood ratio:

$$\log \frac{\mathrm{d}\mathbb{P}_{\theta_0+\frac{t}{\sqrt{n}}}^{\otimes n}}{\mathrm{d}\mathbb{P}_{\theta_0}^{\otimes n}} = t^\top \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}(X_i, \theta_0) - \frac{1}{2} t^\top I(\theta_0) t + o_{\mathbb{P}_{\theta_0}}(1) \rightsquigarrow N\left(-\frac{1}{2} t^\top I(\theta_0) t, t^\top \mathbb{P}_{\theta_0} \dot{\ell}(\theta_0) \dot{\ell}(\theta_0)^\top t\right)$$

where we denote $\ell(x, \theta)$ as the log-likelihood function at the parameter value θ and $\dot{\ell}(\theta)$ is its partial derivative with respect to the argument θ. This is the so-called LAN (local asymptotic normal) expansion of the log-likelihood ratio. A very powerful sufficient condition for LAN expansion to hold is the following.

Definition 52 (Differentiable in quadratic mean (DQM))**.** A statistical model $(\mathbb{P}_\theta, \theta \in \Theta)$ is DQM at $\theta_0 \in \Theta$ if there exists a score $\dot{\ell}(\theta)$ such that the following weak differentiability condition in squared Hellinger distance holds

$$\int \left(\sqrt{\mathrm{d}\mathbb{P}_{\theta_0+h}} - \sqrt{\mathrm{d}\mathbb{P}_{\theta_0}} - \frac{1}{2} h^\top \dot{\ell}(\theta_0) \sqrt{\mathrm{d}\mathbb{P}_{\theta_0}}\right)^2 \mathrm{d}\mu = o\left(\|h\|^2\right). \tag{51}$$

Here the Fisher information is $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\dot{\ell}(\theta_0)\dot{\ell}(\theta_0)^\top\right]$.

Lemma 53. *DQM at θ implies the following:*

1. *$\mathbb{P}_\theta \dot{\ell}(\theta) = 0$ and $I(\theta) = \mathbb{P}_\theta \dot{\ell}(\theta)\dot{\ell}(\theta)^\top$ exists.*
2. *DQM at θ implies LAN expansion at θ.*

Proof. The first part is trivial. We focus on the second part. We abbreviate $\mathrm{d}\mathbb{P}_{\theta_0+\frac{t}{\sqrt{n}}}$ as p_n and $\mathrm{d}\mathbb{P}_{\theta_0}$ as p_0.

$$\begin{aligned}
\log \frac{\mathrm{d}\mathbb{P}_{\theta_0+\frac{t}{\sqrt{n}}}^{\otimes n}}{\mathrm{d}\mathbb{P}_{\theta_0}^{\otimes n}} &= 2\sum_{i=1}^n \log \frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} \\
&= 2\sum_{i=1}^n \log\left(1 + \left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} - 1\right\}\right) \\
&= 2\sum_{i=1}^n \left[\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} - 1\right\} - \frac{1}{2}\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} - 1\right\}^2 + o\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} - 1\right\}^3\right].
\end{aligned}$$

Can you see why the small order term is true? By DQM,

$$\begin{aligned}
&\mathsf{var}_{\theta_0}\left[\sum_{i=1}^n \left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}} - 1\right\} - \frac{1}{2\sqrt{n}} \sum_{i=1}^n t^\top \dot{\ell}(X_i, \theta_0)\right] \\
\le\ & n \int \left(\sqrt{\mathrm{d}\mathbb{P}_{\theta_0+h}} - \sqrt{\mathrm{d}\mathbb{P}_{\theta_0}} - \frac{1}{2} h^\top \dot{\ell}(\theta_0) \sqrt{\mathrm{d}\mathbb{P}_{\theta_0}}\right)^2 = no\left(\frac{\|t\|^2}{n}\right) = o\left(\|t\|^2\right)
\end{aligned}$$

and

$$\mathbb{E}_{\theta_0}\left[2\sum_{i=1}^{n}\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}}-1\right\}\right]=2n\int\sqrt{p_np_0}-1=-n\int(\sqrt{p_n}-\sqrt{p_0})^2\to-\frac{1}{4}\mathbb{P}_{\theta_0}t^\top[\dot\ell(\theta_0)\dot\ell(\theta_0)^\top]t.$$

Thus we have

$$2\sum_{i=1}^{n}\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}}-1\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}t^\top\dot\ell(X_i,\theta_0)-\frac{1}{4}t^\top\mathbb{P}_{\theta_0}[\dot\ell(\theta_0)\dot\ell(\theta_0)^\top]t+o_{\mathbb{P}_{\theta_0}}(1).$$

Next it is easy to see

$$\sum_{i=1}^{n}\left\{\frac{\sqrt{p_n(X_i)}}{\sqrt{p_0(X_i)}}-1\right\}^2\to_{\mathbb{P}_{\theta_0}}\frac{1}{4}t^\top\mathbb{P}_{\theta_0}[\dot\ell(\theta_0)\dot\ell(\theta_0)^\top]t$$

and finally

$$\log\frac{\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_0+\frac{t}{\sqrt{n}}}}{\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_0}}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}t^\top\dot\ell(X_i,\theta_0)-\frac{1}{2}t^\top\mathbb{P}_{\theta_0}[\dot\ell(\theta_0)\dot\ell(\theta_0)^\top]t+o_{\mathbb{P}_{\theta_0}}(1).$$

□

Remark 54. We will see an instance how DQM is used in the paper [**?**] presentation.

6.2 Convolution theorem and Local Asymptotic Minimax (LAM) theorem

Lucien Le Cam spent most of his career trying to develop a general theory for the optimality of MLE by taking the Hodges' estimator into account.

Le Cam's first attempt: around 1953, in his PhD thesis [**?**], Le Cam showed that Hodges' estimator is super-efficient (more efficient than MLE) only on a set of Lebesgue measure zero.

Le Cam's second attempt: restrict only to "(locally) regular estimators T_n".

Definition 55 ((Locally) regular estimators)**.** (Locally) regular estimators T_n of a parameter $\psi(\theta_0), \theta_0 \in \Theta \subset \mathbb{R}^d$ satisfy the following criterion:

$$\sqrt{n}\left\{T_n-\psi\left(\theta_0+\frac{h}{\sqrt{n}}\right)\right\}\rightsquigarrow L(\theta_0)$$

where $L(\theta_0)$ is some tight probability distribution that depends on θ_0, and the convergence is with respect to every law $\mathbb{P}^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}$ for any $h \in \mathbb{R}^d$.

Remark 56. Regular estimators are very restrictive – the asymptotic law of which under local Pitman's alternative should be the same. You can check that Hodges' estimator is not regular. Similarly, Lasso estimator for linear model regression coefficients (β in $y = X\beta +$ noise) is also not regular. To connect with our previous discussion, this is why standard confidence intervals associated with Lasso regression coefficients do not have the correct coverage (essentially a Hodges' estimator). To obtain valid inference, one needs to restore the "regularity" of Lasso estimator by the so-called debiased Lasso (which is correcting for the first-order influence function of the coefficients β).

Theorem 57 (Hajék-Le Cam convolution theorem)**.** *Assuming the model is DQM. For regular estimator T_n, there exists some probability measure $M(\theta)$ such that*

$$L(\theta) = N\left(0, \dot{\psi}(\theta)^\top I(\theta)^{-1} \dot{\psi}(\theta)\right) * M(\theta)$$

where $*$ denotes the convolution between the two distributions.

The proof of the convolution theorem relies on four lemmas by Le Cam (Le Cam's first lemma to Le Cam's fourth lemma; but the most important one is Le Cam's third lemma). Due to time limitation, I am leaning toward skipping the proof. You may find a quick introduction in Jon Wellner's notes or David Pollard's notes or simply read Chapters 6, 8.5 and 8.6 of [**?**]. Of course, you can also read Le Cam's very own treatise [**?**]. I'll give a brief sketch of how Le Cam's program is applied in general: Le Cam's program is trying to show the following: if $X_n \rightsquigarrow_{\mathbb{P}_n} X$ then under a contiguous measure $\mathbb{Q}_n$ (local alternative), $X_n \rightsquigarrow_{\mathbb{P}_n} X'$.

1. Under DQM, which implies LAN, Le Cam's second lemma says

   Lemma 58 (Le Cam's second lemma)**.** *Under DQM, we have*

   $$\log L_n \rightsquigarrow_{\mathbb{P}_n} N(-\frac{\sigma^2}{2}, \sigma^2).$$

   where $L_n \coloneqq \dfrac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}$.

2. Show contiguity (asymptotic absolutely continuity) of the local alternative law $\mathbb{Q}_n$ relative to the null law $\mathbb{P}_n$ (denoted as $\mathbb{Q}_n \triangleleft \mathbb{P}_n$), possibly by Le Cam's first lemma (equivalent characterizations of contiguity):

   Lemma 59 (Le Cam's first lemma)**.** *If $\log L_n \rightsquigarrow_{\mathbb{P}_n} \log L = N(-\frac{\sigma^2}{2}, \sigma^2)$, then $\mathbb{Q}_n \triangleleft \mathbb{P}_n$.*

3. Le Cam's third lemma:

   Lemma 60 (Le Cam's third lemma)**.** *If $\mathbb{Q}_n \triangleleft \mathbb{P}_n$*

   $$\left(X_n, \log \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right) \rightsquigarrow_{\mathbb{P}_n} (X, V),$$

   then the probability measure $L(B) \coloneqq \mathbb{E}\left[\mathbb{1}\{X \in B\} V\right]$ (check on your own why this is a probability measure), then $X_n \rightsquigarrow_{\mathbb{Q}_n} L$.

4. People often use the following corollary of Le Cam's third lemma: If

   $$\left(X_n, \log \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right) \rightsquigarrow_{\mathbb{P}_n} N\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{pmatrix}\right)$$

   then

   $$X_n \rightsquigarrow_{\mathbb{Q}_n} N(\mu + \tau, \Sigma).$$

A recent important paper on independence testing by Fang Han uses this machinery but it is a very standard application (i.e. not in modern high-dimensional or non-parametric settings).

Le Cam's third attempt: Local asymptotic minimax theorem.

Theorem 61 (Local Asymptotic Minimax (LAM) Theorem)**.** *Assuming the model is DQM. For symmetric and convex loss function ℓ[8], then*

$$\inf_{T_n} \lim_{c\to\infty} \liminf_{n\to\infty} \sup_{\|t\|\le c} \mathbb{E}_{\theta+\frac{t}{\sqrt{n}}} \left[\ell\left(\sqrt{n}\left\{T_n - \psi\left(\theta+\frac{t}{\sqrt{n}}\right)\right\}\right)\right] \ge \mathbb{E}[\ell(Z)]$$

where $Z \sim N(0, \dot\psi(\theta)^\top I(\theta)^{-1}\dot\psi(\theta))$.

The proof of LAM also relies on Le Cam's four lemmata. But we do not repeat their statements here.

Proof sketch. The high-level strategy is as follows: DQM $\Rightarrow$ LAN expansion $\Rightarrow$ weak convergence of the likelihood ratio to a Gaussian limiting model $\Rightarrow$ Under the Gaussian limiting model, the risk of MLE cannot be improved in the minimax sense (Anderson lemma; see Lemma 64) by taking a prior $N(0,\Gamma)$ over t and let $\Gamma\to\infty$ in a suitable sense.

The proof technique that we present here is quite useful in many other settings. It is based on "exponential tilting" and "truncating the likelihood ratio statistic". We may see such technique again when we talk about low-degree polynomial methods in "computational-statistical gap" part.

Without loss of generality, we take $\theta_0 = 0$, $\psi = \mathrm{id}$ and obliviate all their appearances in our notation. We denote $I(\theta)^{-1}$ at $\theta = 0$ as Σ. For convenience, we denote $\underline{X}_n = (X_1, \cdots, X_n)^\top$.

Step 1: By LAN and Le Cam's third lemma, we realize that the key part of the likelihood expansion is $Z_n \equiv Z_n(\underline{X}_n) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \Sigma\dot\ell(X_i)$ and $Z_n \rightsquigarrow N(t,\Sigma)$ under $\mathbb{P}_{t/\sqrt{n}}$.

Step 2: Let us try to lower bound the risk directly: denote the law of data given the parameter $t/\sqrt{n}$ as $\mathbb{P}_{t/\sqrt{n},n}$ and consider a truncated Gaussian prior $\Pi_\Gamma^c[t] = N(0,\Gamma)[t]\mathbb{1}\{\|t\|\le c\}$ of t so $\Pi_\Gamma^\infty = N(0,\Gamma)$. With slight abuse of notation, we denote the marginal of $\underline{X}_n$ under the prior Π_Γ^c as $\mathbb{P}_{X,n}$ and the posterior of t as $\Pi_\Gamma^c[t|\underline{X}_n]$.

$$\begin{aligned}
&\sup_{\|t\|\le c} \mathbb{E}_{\frac{t}{\sqrt{n}}}\left[\ell\left(\sqrt{n}\left\{T_n - \frac{t}{\sqrt{n}}\right\}\right)\right] \\
&\ge \int\int \ell\left(\sqrt{n}\left\{T_n - \frac{t}{\sqrt{n}}\right\}\right) \mathrm{d}\mathbb{P}_{t/\sqrt{n},n}(\underline{X}_n)\mathrm{d}\Pi_\Gamma^c[t] \\
&\ge \int \mathbb{E}_{\Pi_\Gamma^c[\cdot|\underline{X}_n]}\ell(\sqrt{n}T_n - t)\mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n) \\
&\ge \int \inf_{\widehat{t}} \mathbb{E}_{\Pi_\Gamma^c[\cdot|\underline{X}_n]}\ell(\widehat{t} - t)\mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n).
\end{aligned}$$

Step 3: Now let us pause for a bit and think about how far we are away from the final statement. In the above lower bound, the joint measure is $\mathrm{d}\mathbb{G}_0(t,\underline{X}_n) \coloneqq \mathrm{d}\mathbb{P}_{t/\sqrt{n},n}(\underline{X}_n)\mathrm{d}\Pi_\Gamma^c[t]$ whereas our target law is $\mathrm{d}\mathbb{G}_\infty(t,\underline{X}_n) \coloneqq \mathrm{d}N(t,\Sigma)[Z_n]\mathrm{d}\Pi_\Gamma^\infty[t]$. One step at a time. Now, exponential tilting (essentially a change of measure).

$$\mathrm{d}\mathbb{G}_0(t,\underline{X}_n) \coloneqq \mathrm{d}\mathbb{P}_{t/\sqrt{n},n}(\underline{X}_n)\mathrm{d}\Pi_\Gamma^c[t] \Rightarrow$$

[8] In most textbooks, bowl-shaped loss is considered – which is symmetric and quasi-convex.

$$\mathrm{d}\mathbb{G}_1(t, \underline{X}_n) \coloneqq \exp\left\{-\frac{1}{2}(Z_n - t)^\top \Sigma^{-1}(Z_n - t)\right\} \underbrace{\exp^{-1}\left\{-\frac{1}{2} Z_n^\top \Sigma^{-1} Z_n\right\} \mathrm{d}\mathbb{P}_{0,n}(\underline{X}_n)}_{\text{likelihood ratio between the null models: finite sample vs. asymptotics}} \mathrm{d}\Pi_\Gamma^c[t]$$

The tilted measure is close to the initial measure simply by LAN.

Step 4: $\mathrm{d}\mathbb{G}_1(t, \underline{X}_n)$ is still different from $\mathrm{d}\mathbb{G}_\infty(t, \underline{X}_n)$. What to do? Change the measure slightly again!

$$\mathrm{d}\mathbb{G}_1(t, \underline{X}_n) \Rightarrow \mathrm{d}\mathbb{G}_2(t, \underline{X}_n) \coloneqq \underbrace{\exp\left\{-\frac{1}{2}(Z_n - t)^\top \Sigma^{-1}(Z_n - t)\right\} \exp^{-1}\left\{-\frac{1}{2} Z_n^\top \Sigma^{-1} Z_n\right\} \mathrm{d}\mathbb{P}_{0,n}(\underline{X}_n)}_{=:\mathrm{d}\mathbb{Q}_{t,n}(\underline{X}_n)} \mathrm{d}\Pi_\Gamma^\infty[t]$$

which is close to the target measure $\mathrm{d}\mathbb{G}_\infty$.

Now we need to show $\mathbb{G}_0 \approx \mathbb{G}_1 \approx \mathbb{G}_2 \approx \mathbb{G}_\infty$. $\mathbb{G}_1$ vs. $\mathbb{G}_2$ seems to be the simplest: we only let $c \to \infty$:

$$\begin{aligned}
&\|\mathbb{G}_1 - \mathbb{G}_2\|_{\mathrm{TV}} \\
\leq &\int \left|\mathrm{d}\mathbb{Q}_{t,n}(\underline{X}_n)\left(\mathrm{d}\Pi_\Gamma^c(t) - \mathrm{d}\Pi_\Gamma^\infty(t)\right)\right| \\
\leq &\int \sup_t \left|\mathrm{d}\mathbb{Q}_{t,n}(\underline{X}_n)\right| \int \left|\mathrm{d}\Pi_\Gamma^c(t) - \mathrm{d}\Pi_\Gamma^\infty(t)\right| \\
\equiv &\int \exp\left\{\frac{1}{2} Z_n^\top \Sigma^{-1} Z_n\right\} \mathrm{d}\mathbb{P}_{0,n}(\underline{X}_n) \|\Pi_\Gamma^c - \Pi_\Gamma^\infty\|_{\mathrm{TV}}.
\end{aligned}$$

Darn! How to bound $\int \exp\left\{\frac{1}{2} Z_n^\top \Sigma^{-1} Z_n\right\} \mathrm{d}\mathbb{P}_{0,n}(\underline{X}_n)$? Now comes the "truncated likelihood ratio" trick.

Step 5: Define a new measure $\bar{\mathbb{P}}_{t/\sqrt{n},n}$ by truncating $\mathbb{P}_{t/\sqrt{n},n}$ within the

$$\mathrm{box} \coloneqq \{\underline{X}_n : \|Z_n(\underline{X}_n)\| \leq b\}.$$

For any truncation error threshold $\epsilon > 0$, we can choose b appropriately so that the difference between $\bar{\mathbb{P}}_{t/\sqrt{n},n}$ and $\mathbb{P}_{t/\sqrt{n},n}$ is bounded by ϵ. The corresponding $\|\bar{\mathbb{G}}_1 - \bar{\mathbb{G}}_2\|_{\mathrm{TV}}$ can be easily shown to be small. Between $\mathrm{d}\bar{\mathbb{G}}_0$ and $\mathrm{d}\bar{\mathbb{G}}_1$, we can show, by LAN expansion,

$$\begin{aligned}
&\liminf_n \sup_{t:\|t\| \leq C} \|\mathbb{Q}_{t,n} - \bar{\mathbb{P}}_{t/\sqrt{n},n}\|_{\mathrm{TV}} \\
= &\liminf_n \sup_{t:\|t\| \leq C} \int_{\mathrm{box}} \left|\frac{\mathrm{d}\bar{\mathbb{Q}}_{t,n}}{\mathrm{d}\bar{\mathbb{P}}_{0,n}} - \frac{\mathrm{d}\bar{\mathbb{P}}_{t/\sqrt{n},n}}{\mathrm{d}\bar{\mathbb{P}}_{0,n}}\right| \mathrm{d}\bar{\mathbb{P}}_{0,n} \\
= &\liminf_n \sup_{t:\|t\| \leq C} \int_{\mathrm{box}} \left|\exp\left\{t^\top \Sigma^{-1} Z_n - \frac{1}{2} t^\top \Sigma^{-1} t\right\}\left(1 - e^{o_{\mathbb{P}_{0,n}}(\|t\|)}\right)\right| \mathrm{d}\bar{\mathbb{P}}_{0,n} \to 0.
\end{aligned}$$

Finally, we go back to the risk lower bound we have left for a while. Denote $N_{\Sigma,\Gamma}$ as the posterior of the likelihood $Z_n \sim N(t, \Sigma)$ and the prior $N(0, \Gamma)$.

$$\int \inf_{\widehat{t}} \mathbb{E}_{\Pi_\Gamma^c[\cdot|\underline{X}_n]} \ell(\widehat{t} - t) \mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n)$$

$$
\begin{aligned}
&= \int \inf_{\widehat{t}} \mathbb{E}_{N_{\Sigma,\Gamma}[\cdot|Z_n]} \ell(\widehat{t}-t) + \mathbb{E}_{\Pi^c_\Gamma[\cdot|\underline{X}_n]} \ell(\widehat{t}-t) - \mathbb{E}_{N_{\Sigma,\Gamma}[\cdot|Z_n]} \ell(\widehat{t}-t) \mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n) \\
&\geq \int \inf_{\widehat{t}} \mathbb{E}_{N_{\Sigma,\Gamma}[\cdot|Z_n]} \ell(\widehat{t}-t) \mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n) - \sup_{\widehat{t},t} \ell(\widehat{t}-t) \int \|N_{\Sigma,\Gamma}(\cdot|Z_n) - \Pi^c_\Gamma[\cdot|\underline{X}_n]\|_{\mathrm{TV}} \mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n).
\end{aligned}
$$

The first term is handled by Anderson lemma. We need to show the second term can be made smaller than any given threshold $\varepsilon > 0$. Denote $\mathbb{P}_{X,n,2}$ as the marginal distribution of $\underline{X}_n$ under the law $\mathrm{d}\mathbb{G}_2$

$$
\begin{aligned}
&\int \|N_{\Sigma,\Gamma}(\cdot|Z_n) - \Pi^c_\Gamma[\cdot|\underline{X}_n]\|_{\mathrm{TV}} \mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n) \\
\leq &\int \|N_{\Sigma,\Gamma}(\cdot|Z_n) - \Pi^c_\Gamma[\cdot|\underline{X}_n]\|_{\mathrm{TV}} \left(\mathrm{d}\mathbb{P}_{X,n}(\underline{X}_n) + \mathrm{d}\mathbb{P}_{X,n,2}(\underline{X}_n)\right) \\
\overset{?}{\leq} &\, 4\|\mathbb{G}_0 - \mathbb{G}_2\|_{\mathrm{TV}}.
\end{aligned}
$$

□

Remark 62. Proof of the final "?".

Lemma 63.

$$
\int \|M_1(\cdot|\underline{X}_n) - M_2(\cdot|\underline{X}_n)\|_{TV} (\mathrm{d}\mu_1(\underline{X}_n) + \mathrm{d}\mu_2(\underline{X}_n)) \leq 4\| \underbrace{M_1 - M_2}_{Joint\ measures} \|_{TV}.
$$

Proof. Denote $\mathrm{d}M_1(\cdot|\underline{X}_n) = a_1$, $\mathrm{d}\mu_1(\underline{X}_n) = b_1$, $\mathrm{d}M_2(\cdot|\underline{X}_n) = a_2$, $\mathrm{d}\mu_2(\underline{X}_n) = b_2$.

Then

$$
\begin{aligned}
&\int\int |a_1 - a_2|(b_1 + b_2) \\
=&\int\int |a_1 - a_2| b_1 + |a_1 - a_2| b_2 \\
=&\int\int |a_1 b_1 - a_2 b_2 + a_2(b_2 - b_1)| + |a_1 b_1 - a_2 b_2 + a_1(b_2 - b_1)| \\
\leq&\int\int 2|a_1 b_1 - a_2 b_2| + (a_1 + a_2)|b_2 - b_1|.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\int \|M_1(\cdot|\underline{X}_n) - M_2(\cdot|\underline{X}_n)\|_{\mathrm{TV}} (\mathrm{d}\mu_1(\underline{X}_n) + \mathrm{d}\mu_2(\underline{X}_n)) \\
\leq &\int 2|\mathrm{d}\mu_1(\underline{X}_n)) \mathrm{d}M_1(h|\underline{X}_n)) - \mathrm{d}\mu_2(\underline{X}_n) \mathrm{d}M_2(h|\underline{X}_n))| \\
&+ \int \{\mathrm{d}M_1(h|\underline{X}_n) + \mathrm{d}M_2(h|\underline{X}_n)\} |\mathrm{d}\mu_1(\underline{X}_n) - \mathrm{d}\mu_2(\underline{X}_n)| \\
\leq &\, 2\|M_1 - M_2\|_{\mathrm{TV}} + 2\|\mu_1 - \mu_2\|_{\mathrm{TV}}.
\end{aligned}
$$

Finally

$$
\int |\mathrm{d}\mu_1 - \mathrm{d}\mu_2| = \int_{\mathbb{X}_n} \left| \int_h \mathrm{d}M_1(h|\underline{X}_n) - \mathrm{d}M_2(h|\underline{X}_n) \right|
$$

$$\leq \int_{\mathbb{X}_n} \int_h |\mathrm{d}M_1 - \mathrm{d}M_2| \leq \|M_1 - M_2\|_{\mathsf{TV}}.$$

□

Lemma 64 (Anderson lemma)**.** *For statistical model $X \sim N(\theta, \Sigma)$, symmetric and convex loss function ℓ,*

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta \left[\ell \left(\widehat{\theta}_n - \theta \right) \right] \geq \mathbb{E} \left[\ell(Z) \right] \tag{52}$$

where $Z \sim N(0, \Sigma)$.

Bernstein-von Mises theorem is the Bayesian analogues of the theory of MLE. We state the theorem below without proof.

Theorem 65 (Bernstein-von Mises theorem)**.** *For a probability model $(\mathbb{P}_\theta, \theta \in \Theta)$ that is DQM at θ_0 and $I(\theta_0)$ is symmetric and positive definite. If the prior Π_θ is absolutely continuous around a neighborhood of θ_0 and has non-zero density at θ_0, then the posterior distribution $\theta | X_1, \cdots, X_n$ converges weakly to $N(\theta_0, I(\theta_0)^{-1})$.*

We may come back to this theorem if we have time to cover non-parametric Bayesian statistics.

In this lecture, I deliberately try to avoid using the name "statistical experiments", which is strongly advocated by mathematical statisticians following Lucien Le Cam's path. This is only because I do not have enough time to tell you what is "statistical experiments" and why "statistical experiments" is an important concept.

Finally, let us return to the start of this chapter: van Tree's inequality. After you have learnt about all those results on MLE, how would you interpret van Tree's inequality?