300 Quals Guide

July 14, 2025

1 Stochastic Convergence

Helpful tools

1. Don't forget Portmanteau
2. CLT, SLLN
3. CMT
4. Slutsky

Theorem 1 (CMT)
If g is continuous on a set of probability 1

$$X_n \to^* X \implies g(X_n) \to^* g(X)$$

for conv in dist, prob, or as

Theorem 2 (Slutsky)
A few parts

1. $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$
2. If $\|X_n - Y_n\| \xrightarrow{p} 0$ then if $X_n \xrightarrow{d} X$ we have also $Y_n \xrightarrow{d} X$.
3. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ then
$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

Corrolary via Slutsky and CMT: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ gives us

1. $Y_n X_n \xrightarrow{d} cX$
2. $Y_n + X_n \xrightarrow{d} c + X$
3. $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$

Definition 3 (Uniform Tightness)
A collection $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight if for all $\epsilon > 0$ there exists an $M < \infty$ such that

$$P(\|X_\alpha\| > M) \leqslant \epsilon \quad \text{for all } \alpha \in A$$

Example 4 (Markov, Tightness)
If all X_α have the same ℓ-th moment, then just use Markov to prove tightness.

In general if a collection has increasing means or something, it won't be tight, eg $X_n \sim N(n, 1)$.

Theorem 5 (Prohorov)
Uniformly tight $\implies$ for all sequences there is a a subsequence that converges in distribution to a random variable.

Conversely, Convergence in distribution $\implies$ unifromly tight.

1.1 Big-O, Little-o

Non-stochastic versions,

$$f(x) = O(g(x)) \iff \limsup_{\varepsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} < \infty$$

.

$$f(x) = o(g(x)) \iff \lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{g(\varepsilon)} = 0$$

Definition 6 (Little-o_p)

$$X_n = o_p(R_n) \iff \exists Y_n \text{ such that } X_n = R_n Y_n \text{ with } Y_n \xrightarrow{p} 0$$

Analogous to how $f(n) = o(n)$ if $f(n) = no(1)$.

Definition 7 (Big O_p)

$$X_n = O_p(R_n) \iff X_n = R_n Y_n \text{ where } Y_n \text{ uniformly tight}$$

Combining o_p and O_p algebra:

1. ....

1.2 Delta Method

Theorem 8 (Delta Method)

Let $r_n \to \infty$ and f differentiable at θ. If $r_n(T_n - \theta) \xrightarrow{d} A$ then:

1. $$r_n(f(T_n) - f(\theta)) \xrightarrow{d} f'_\theta A$$

2. $$r_n(f(t_n) - f(\theta)) - f'_\theta(r_n(T_n - \theta)) \xrightarrow{p} 0$$

Proof. Main idea:

$$f(\theta + h) - f(\theta) = f'_\theta h + o(\|h\|) \text{ as } h \to 0$$

Take $h = T_n - \theta$ □

Theorem 9 (Higher order delta method)

Example 10 (...)

2 MLE

Definition 11 (M-estimators)

$$\hat{\theta}_n = \arg\max_\theta \frac{1}{n} \sum_{i=1}^n m_\theta(X_i)$$

where m_θ is some known function, eg ℓ_θ– maximize log likelihood

Definition 12 (Z-estimator)

$$\hat{\theta}_n = \{\theta : n^{-1} \sum_{i=1}^n \Psi_\theta(X_i) = 0\}$$

eg $\nabla \ell_\theta = 0$.

2.1 Consistency of MLE

Consistency of $P_n \ell_\theta \xrightarrow{p} P\ell_\theta$ (WLLN) for any θ fixed is **not** enough for consistency of MLE $\hat{\theta}_n \xrightarrow{p} \theta^*$. For arbitrary sample size, $P_n \ell_\theta$ not necessarily maximized at (or near) θ^* if we converge too non-uniformly. See picture. For consistency of MLE, need:

1. Uniform convergence (in probability) ie

$$\sup_{\theta \in \Theta} |P_n \ell(\theta) - P\ell(\theta)| = o_p(1)$$

2. Well-separation

Definition 13 (Uniform convergence)
$M_n(\theta)$ converges uniformly to $M(\theta)$ if

$$\sup_\theta |M_n(\theta - M(\theta)| \xrightarrow{p} 0$$

Definition 14 (Well-separation)
definition.....
eg, strong convexity.

Some more primitive conditions. Identifiability and a finite sample space $|\Theta| < \infty$ are enough for consistency of MLE.

Definition 15 (Identifiability)
Identifiable if for $\theta \neq \theta'$, $P_\theta \neq P'_\theta$ ie KL divergence is strictly positive.

Example 16 (Example of non-identifiability)
In my notes

2.2 Asymptotic normality of the MLE

Under a regularity condition ("smooth, nice"), the MLE is normal.

Definition 17 (Smooth/Nice at θ)
See notes..

1. Hessian of log likelihood is Lipschitz near θ^* - see notes
2. Bounded gradient $P^*_\theta \|\nabla \ell^*_\theta\|^2 < \infty$

Theorem 18 (Asymptotic normality of MLE)
If

1. $\{P_\theta\}_{\theta \in \Theta}$ is smooth/nice at θ^*
2. Θ open subset of $\mathbb{R}^d$,
3. Hessian has finite mean (or alt that exchange order of differentiation wrt θ and expectation).
4. $\hat{\theta}_n$ the MLE is consistent

Then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma^*_\theta)$. **If we can also exchange order of differentiation and expectation** , $\Sigma^*_\theta = I_\theta^{-1}$.

3 Fisher Information

Definition 19 (Fisher Information)
Outer product of the score

$$I_\theta = \mathbf{E}_{P_\theta}[\nabla_\theta \ell_\theta (\nabla_\theta \ell_\theta)^T]$$

If we can exchange order of differentiation and expectation then:

$$I_\theta = \mathbf{Cov}(\nabla \ell_\theta) = -\mathbf{E}[\nabla^2 \ell_\theta]$$

Lots of information, small variance. Look at hessian to look at curvature– if it's really peaked, we have lots of information.

Theorem 20 (Cramer-Rao)
If δ_n is unbiased for $\theta \in \mathbb{R}^p$, then

$$\mathbf{Cov}_\theta(\delta_n) = \mathbf{E}_\theta[(\theta - \delta_n)(\theta - \delta_n)^T] \succeq \frac{1}{n} I_\theta^{-1}$$

Distribution	**Parameter(s)**	**Fisher information** $I(\theta)$
Bernoulli Bern(p)	$p \in (0,1)$	$\frac{1}{p(1-p)}$
Binomial Bin(m, p) (fixed m)	$p \in (0,1)$	$\frac{m}{p(1-p)}$
Poisson Pois(λ)	$\lambda > 0$	$\frac{1}{\lambda}$
Exponential Exp(λ)	$\lambda > 0$	$\frac{1}{\lambda^2}$
Gamma Gamma(α, θ) (fixed α)	$\theta > 0$	$\frac{\alpha}{\theta^2}$
Normal $\mathcal{N}(\mu, \sigma^2)$ (known σ^2)	$\mu \in \mathbb{R}$	$\frac{1}{\sigma^2}$
Normal $\mathcal{N}(\mu, \sigma^2)$ (known μ)	$\sigma^2 > 0$	$\frac{1}{2\sigma^4}$
Normal $\mathcal{N}(\mu, \sigma^2)$ (both unknown)		$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$

Definition 21 (Asymptotic efficiency)
T_n is efficient *for* θ if

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, I_\theta^{-1})$$

4 Efficiency

Efficient estimators:

1. MLE for nice families
2. MoM estimator in some exponential families
3. One-step estimator from the homework

5 Asymptotic hypothesis testing

Definition 22 (Uniform asymptotic level)
The more interesting level to look at.

$$\limsup_n \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(\text{Test } T_n \text{ rejects null})$$

Intuitively, the limit of the typical "size" of the test.

Definition 23 (Pointwise asymptotic level)

$$\sup_{\theta_0 \in \Theta_0} \limsup_n P_{\theta_0}(T_n \text{ rejects null})$$

Pointwise limit, then take sup.

6 Contiguity

Beginning discussion of optimality of an estimator. Want to build tools to zoom in on the interesting scale. The motivation is that if we fix $\theta' \neq \theta^*$, if our estimator is asymptotically normal for example, we can eventually tell these two apart. Ie, we can eventually reject $\mathcal{H}_0 : \theta = \theta'$. Ie if $\|\theta' - \theta^*\| \geqslant \delta$, we have:

$$\lim_n P_{\theta^*}(\|\hat{\theta}_n - \theta'\| \geqslant 0.9\delta) = 1$$

Conversely, if $\|\theta' - \theta^*\| = o(1/\sqrt{n})$, it will be impossible to distinguish θ' and θ^*.

Example 24 (See HW 4 example pointwise versus uniform level α...)

This motivates $\theta_n = \theta_0 + h/\sqrt{n}$ where $\|h\| = O(1)$ and testing:

$$\mathcal{H}_0 : \theta^* = \theta_0 \quad \text{vs} \quad \mathcal{H}_1 : \theta^* = \theta_n$$

Want to know the power of a test at this scale and How good $\hat{\theta}_n$ can be if it must be good under θ_0 and under θ_n to rule out adversarial examples like Hodge's?

Definition 25 (Contiguity)
Let P_n, Q_n be sequences of probability measures on the same sample space for each n. $Q_n \triangleleft P_n$ means Q_n is contiguous wrt P_n if for any sequence of events A_n:

$$P_n(A_n) \to 0 \implies Q_n(A_n) \to 0.$$

Mutually contiguous makes the above $\iff$.

Example 26 (Example of $P_n \triangleleft Q_n$ but not $Q_n \triangleleft P_n$)
Just let $Q_n = N(0,1), P_n = |N(0,1)|$ works since $P_n \ll Q_n$.

6.1 Le Cam's Lemmas

Let P_n, Q_n be dominated by μ_n such that have densities p_n, q_n. Define L_n likelihood ratio.

Theorem 27 (Le Cam 1)
The following are equivalent:

1. $Q_n \triangleleft P_n$
2. If $\frac{1}{L_n} \xrightarrow[Q_n]{d} U$ on a subsequence, then $P(U > 0) = 1$
3. If $L_n \xrightarrow[P_n]{d} L$ on a subsequence, then $\mathbf{E}L = 1$.
4. If $T_n \xrightarrow[P_n]{p} 0$, then $T_n \xrightarrow[Q_n]{p} 0$

Main idea is that the limiting behavior of the likelihood ratio determines whether we have contiguity.

Example 28 (Contiguity of the log likelihood)
If $\log L_n \xrightarrow[Q_n]{d} N(\mu, \sigma^2)$, then by CMT and condition 2 of the above we get that $Q_n \triangleleft P_n$. If also $\mu = -\frac{1}{2}\sigma^2$, we get $P_n \triangleleft Q_n$.

We have this exact situation if P_θ allows under some "niceness" –then the log likelihood ratio of $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$ looking at the log likelihood ratio of the local alternative versus null. So the local alternative measure and the P_{θ_0} measure are mutually contiguous in the iid case.

Theorem 29 (Le Cam's Third)
If $Q_n \triangleleft P_n$ and $(X_n, L_n) \xrightarrow[P_n]{d} (X, L)$ then:

$$(X_n, L_n) \xrightarrow[Q_n]{d} W \quad \text{where } (X, L) \overset{D}{=} M \text{ and } dW(x, \ell) = \ell \, dM(x, \ell)$$

Other stuff..

Theorem 30 (The actual useful Le Cam Corrollary)
If

$$(X_n, \log L_n) \xrightarrow[P_n]{d} (X, Z) \sim N\left(\begin{pmatrix} \mu \\ \frac{-1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right),$$

then we have

$$X_n \xrightarrow[Q_n]{d} N(\mu + \tau, \Sigma).$$

"If we know the limiting distribution of X_n and the likelihood ratio under P_{θ_0}, then we know that X_n is also normal under the local alternative. We can transfer information from P_{θ_0} to $P_{\theta_0 + \frac{1}{\sqrt{n}}h}$."

The point of the above is that we want to understand the power of tests in the local regime– to do so we need to be able to analyze the limiting distribution of some statistic under the local alternative.

7 Tail bounds, Subexponential R.V.s

First tools are Markov and its relatives. To get the best bound of this sort, if the mgf exists:

$$P[(X-\mu) \geqslant t] \leqslant \frac{\mathbf{E}[\exp(\lambda(X-\mu))]}{\exp(\lambda t)}, \quad \text{then optimize over } \lambda$$

We can classify r.v.s interms of their mgfs, leading to subgaussian definition. For a normal rv,

$$\mathbf{E}[\exp(\lambda X)] = \exp(\mu\lambda + \sigma^2\lambda^2/2).$$

Note that $\inf_{\lambda \geqslant 0} \log \mathbf{E}[\exp(\lambda(X-\mu)] - \lambda t = -\frac{t^2}{2\sigma^2}$, so we get

$$\boxed{P(X-\mu \geqslant t) \leqslant \exp(-\frac{t^2}{2\sigma^2})} \quad \text{for } X \sim N(\mu, \sigma^2)$$

Definition 31 (Sub-gaussian rv)
X is subgaussian if there exists σ such that:

$$\mathbf{E}[\exp(\lambda(X-\mu))] \leqslant \exp(\sigma^2\lambda^2/2) \quad \forall \lambda \in \mathbb{R},$$

compare to the mgf of normal– asking that the mgf be bounded by that of normal with some σ^2.

Combined with the Chernoff bound:

Theorem 32 (Sub-gaussian concentration inequality)
If X is σ-sub-gaussian:

$$P(|X-\mu| \geqslant t) \leqslant 2\exp(-\frac{t^2}{2\sigma^2}).$$

Example 33 (Bounded rvs)
If $\text{supp}(X) \subseteq [a,b]$, then X is $\frac{b-a}{2}$ subgaussian.

Fact 34 (Subgaussianity preserved by linear ops)
X_1, X_2 are σ_1, σ_2 subgaussian, then $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$ sub Guassian.

An immediate consequence of the above:

Theorem 35 (Hoeffding Bound on concentration of sums of subgaussian rvs)
If X_i are independent with mean μ_i and subgaussian parameter σ_i, then for all $t \geqslant 0$:

$$P(\sum_{i=1}^n (X_i - \mu_i) \geqslant t) \leqslant \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}).$$

If $X_i \in [a,b]$ for all i, then the rhs is $\exp(-\frac{2t^2}{n(b-a)^2})$.

Fact 36 (Many faces of sub-Gaussians)
See HW and Wainwright 2.6.

1. X σ-subgaussian then cX is $|c|\sigma-$ subgaussian.
2. Converse of the chernoff bound: if $P(|X - \mathbf{E}X| \geqslant t) \leqslant 2\exp(-\frac{1}{2\sigma^2}t^2)$ for all $t \geqslant 0$, then X is $C\sigma$-subgaussian.
3. If X is σ subgaussian, then X^2 is $(C_1\sigma^2, C_2\sigma^2)$ subexponential.

Definition 37 (Sub-exponential random variable)
A relaxation of sub-gaussian. X is σ, α sub-exponential if:

$$\mathbf{E}[\exp(\lambda(X-\mu))] \leqslant \exp(\sigma^2\lambda^2/2) \quad \forall|\lambda| < \frac{1}{\alpha}.$$

Ie subgaussian bound, but only for small enough λ.

Subgaussian $\implies$ subexponential but not the converse.

Fact 38 (Many faces of sub-exponential..)

Example 39 (χ^2_n Subexponential)
with parameters $(2\sqrt{n}, 4)$. But not subgaussian!

Just as with sub-gaussian, sum of sub-exponential random variables is subexponential.

Theorem 40 (Sub-exponential concentration)
..

7.1 Martingale Bounds (AH)

Important example is sum of iid random variables of mean 0.

Theorem 41 (Martingale Concentration Inequality (Azuma Hoeffding))
If $Y_0, \ldots,$ is a martingale sequence with conditional martingale differences $(Y_{i+1} - Y_i)|\mathcal{F}_i$ σ_{i+1}-subgaussian, then for all $t > 0$:

$$\boxed{P(|Y_n - Y_0| > t) \leqslant 2\exp(-t^2/(2\sum_{i=1}^{n}\sigma_i^2))}$$

The hypothesis is the same as

$$\mathbf{E}[\exp(\lambda D_k)|\mathcal{F}_{k-1}] \leqslant \exp(\lambda^2\sigma_k^2/2)$$

ie differences are conditionally subgaussian.

And there's a version for subexponential increments as well.

Definition 42 (Bounded differences property)
If x, x' only differ in coordinate k,

$$|g(x) - g(x')| \leqslant L_k.$$

Theorem 43 (Bounded differences inequality)
If f satisfies bounded differences, and $\underline{X}$ has independent components, then

$$P(|f(X) - \mathbf{E}[f(X)]| \geqslant t) \leqslant 2\exp(-\frac{2t^2}{\sum_{k=1}^n L_k^2}).$$

The idea is to use a Doob martingale $Y_k = \mathbf{E}[f(X)|X_1, \ldots, X_k]$ and Azuma Hoeffding (ie in the case of bounded random variables– the random variable $D_k = Y_k - Y_{k-1}$ is in an interval of length L_k. Apply AH for bounded mg differences- Wainwright 2.20).

8 Uniform LLN

Why do we care about uniform convergence of the empirical CDF? Suppose we have a functional γ that maps $F \mapsto \gamma(F)$. For example, $\gamma_g(F) = \mathbf{E}_F g(X)$ is a functional. Another example is quantile functional. If we can show that γ is continuous with respect to the sup norm, then for any ϵ, there exists a δ such that:

$$P(|\gamma(\hat{F}_n) - \gamma(F)| > \epsilon) \leqslant P(\sup_\theta |\hat{F}_n - F| > \delta) \longrightarrow 0.$$

So uniform convergence of the empirical CDF gives us consistency of our estimator $\gamma(\hat{F}_n)$ for $\gamma(F)$. If we have almost sure convergence as below, then also:

$$1 = P(\sup_\theta |\hat{F}_n(\theta) - F(\theta)| \longrightarrow 0) \leqslant P(\lim \gamma(\hat{F}_n) = \gamma(F)),$$

so we get almost sure convergence of our estimator.

Theorem 44 (Glivenko-Cantelli)

$$\sup_{\theta \in \Theta} |\hat{F}_n(\theta) - F(\theta)| \xrightarrow{\text{a.s.}} 0$$

8.1 ULLN for General Function Classes

Back up and try to understand ULLN for more general setup. Define:

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbf{E}[f(X)] \right|.$$

Similarity to empirical cdf setup with $f_\theta(X_i) = \mathbf{1}[X_i \leqslant \theta]$, $\mathcal{F} = \{f_\theta, \theta \in \mathbb{R}\}$. We know from LLN that for each f, $P_n f \xrightarrow{p} Pf$, but when can we say more that $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0$, ie uniform convergence?

Definition 45 (Glivenko-Cantelli Class)
$\mathcal{F}$ is a GC class if $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0$.

Eg class of indicators is a Glivenko-Cantelli class.

If $\hat{\theta}_n$ minimizes the empirical risk, ie:

$$\hat{\theta}_n = \arg\min_{\theta} L_n(\theta) = \arg\min_{\theta} n^{-1} \sum_{i=1}^{n} L_\theta(X_i)$$

eg find me the best $\hat{\beta}$ to minimize $L_n(\beta) = n^{-1}\|X\beta - y\|^2$.

Definition 46 (Excess Risk)
With $L(\theta) = \mathbf{E}_{\theta^*}[L_\theta(X)]$,

$$\text{excess risk} = L(\hat{\theta}_n) - L(\theta^*) = \mathbf{E}_{\theta^*} L(\hat{\theta}_n, X) - \mathbf{E}_{\theta^*} L(\theta^*, X)$$

ie *in expectation*, how much worse is our estimator at the task than the true value? (Or if the true value is not in the class we're considering, how much worse is our estimator than the best in class value?)

Example 47 (MLE)
If we take $L_\theta(x) = -\ell_\theta(x)$, then whether or not $\mathcal{F} = \{\ell_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli tells us whether or not we have uniform convergence, a necessary condition for consistency of MLE.

The excess risk in this case is: $\mathbf{E}_{\theta^*}[\frac{\log p_{\theta^*}}{\log p_{\hat{\theta}_n}}] = KL(p(\theta^*)\|p(\hat{\theta}_n))$.

Other reasonable loss functions:

1. Binary classification: $L_f(X, Y) = \mathbf{1}[f(X) \neq Y]$.
2. Linear regression: $L_\beta(X, Y) = \|Y - X\beta\|^2$

Fact 48 (Controlling excess risk)
How do we control excess risk? By *controlling the generalization gap*, ie

$$\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| = o_p(1) \implies L(\hat{\theta}_n) \xrightarrow{p} L(\theta^*),$$

ie the excess risk going to 0 in probability. That is, $\mathbf{E}_{\theta^*}[L(\hat{\theta}_n, X)] \xrightarrow{p} \mathbf{E}_{\theta^*}[L(\theta^*, X)]$.

Furthermore,

$$\text{excess risk} = L(\hat{\theta}_n) - L(\theta^*) \leqslant 2\|P_n - P\|_{\mathcal{L}(\Omega_0)}.$$

So we control the rhs to control the excess risk.

Thus what we want is a ULLN for the class $\mathcal{L}(\Omega_0) = \{x \mapsto L_\theta(X), \theta \in \Omega_0\}$ since this will give us excess risk $o_p(1)$. So we start to introduce tools to get ULLN (ie bound the probability of extreme values of $\|P_n f - Pf\|_{\mathcal{F}}$ to then get bounds on excess risk when we choose $\mathcal{F}$ as above, ie $f(X) = L_\theta(X)$.

9 Rademacher Complexity

Think of Rademacher Complexity of answering: "how well can our class f correlate with random signs ϵ_i? If we can always find an f that correlates strongly, then our class of functions must be really big. If our class is smaller, then it's easier to bound Rademacher complexity.

Definition 49 (Empirical Rademacher Complexity)

$$\mathcal{R}(\mathcal{F}, x) = \mathbf{E}_\epsilon \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \quad \text{where } x \text{ is fixed, expectation is over random signs only}$$

Definition 50 (Rademacher Complexity)
The expectation of the empirical rademacher complexity.

$$\mathcal{R}_n(\mathcal{F}) := \mathbf{E}_X[\mathcal{R}(\mathcal{F}, X)] = \mathbf{E}_{X,\epsilon} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|$$

We show that controlling the Rademacher Complexity uniformly controls the generalization gap, $\|P_n f - Pf\|_{\mathcal{F}}$, which controls the excess risk.

Theorem 51 (b-uniformly bounded class (Wainwright Prop 4.11))
Any $n \geqslant 1$, $\delta \geqslant 0$, b-uniformly bounded class $\mathcal{F}$:

$$\|P_n - P\|_{\mathcal{F}} \leqslant 2\mathcal{R}_n(\mathcal{F}) + \delta, \qquad \text{w.p. } \geqslant 1 - \exp(-\frac{n\delta^2}{2b^2}).$$

That is, controlling the $\mathcal{R}_n$ allows us to get a bound on the probability of an extreme deviation of $\|P_n - P\|_{\mathcal{F}}$.

Proof. Uses the bounded differences method applied to $G(x) = \sup_f \left| n^{-1} \sum_{i=1}^{n} \overline{f}(x_i) \right|$, so that deviations from $\mathbf{E}[\|P_n - P\|_{\mathcal{F}}]$ happen with probability $\geqslant 1 - \exp(-\frac{n\delta^2}{2b^2})$.

Then we show that $\mathbf{E}[\|P_n - P\|_{\mathcal{F}}] \leqslant 2\mathcal{R}_n(\mathcal{F})$ via symmetrization and arguing that $\epsilon(f(X_i) - f(Y_i)) \overset{d}{=} f(X_i) - f(Y_i)$. □

Theorem 52 (Matching lower bound (Wainwright 4.12))
For same setup as previous theorem,

$$P_n - P\|_{\mathcal{F}} \geqslant \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{1}{2\sqrt{n}} \sup_f |\mathbf{E}[f]| - \delta.$$

from this we can conclude that Rademacher Complexity necessarily must go to 0 to get $\|P_n - P\|_{\mathcal{F}} = o_p(1)$. Necessary and sufficient!

Tools to bound the Rademacher Complexity:

1. Union bound (eg finite function classes)
2. Polynomial discrimination (eg Glivenko-Cantelli result)

(a) Directly as in Glivenko-Cantelli

(b) VC Dimension

3. Metric entropy

4. Todo this later after reviewing

9.1 Polynomial Discrimination

First define:

Definition 53 ($\mathcal{F}(x_1, \dots, x_n)$)

$$\mathcal{F}(x_1, \dots, x_n) = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}$$

ie, what points in $\mathbb{R}^n$ can you hit with the functions in your class and the data you're given.

The cardinality $|\mathcal{F}(x)|$ provides a "sample dependent" idea of the complexity of $\mathcal{F}$. This cardinality only is helpful when it's finite. Eg, if $\mathcal{F}$ is a collection of classifiers $x \mapsto \{0, 1\}$, then $|\mathcal{F}(x_1, \dots, x_n)| \leqslant 2^n$. This is the case with $\mathcal{F} = \{\mathbf{1}[x < \theta] : \theta \in \mathbb{R}\}$, ie the setup for Glivenko Cantelli.

Definition 54 (Polynomial discrimination)
$\mathcal{F}$ has polynomial discrimination of order $v \geqslant 1$ if: for all $n \in \mathbb{Z}^+$, sample $x_1, \dots, x_n$ of n points, we have

$$|\mathcal{F}(x_1, \dots, x_n)| \leqslant (n+1)^v$$

.

To show polynomial discrimination directly:

1. Let n be arbitrary.

2. Let $x_1, \dots, x_n \in \mathcal{X}$ be arbitrary samples.

3. Show the bound.

Theorem 55 (Polynomial discrimination controls rademacher complexity)

$$\mathcal{R}(\mathcal{F}(x)) \leqslant 4 \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}} \sqrt{\frac{v \log(n+1)}{n}}$$

where the LHS is the *empirical* Rademacher complexity (ie a random variable).

As a corrolary, if we can get a bound on $\mathbf{E}_X[\sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}]$, ie if the function class is uniformly b-bounded, then:

$$\mathcal{R}_n(\mathcal{F}) \leqslant 2b \sqrt{\frac{v \log(n+1)}{n}} \quad \forall n \geqslant 1.$$

Fact 56 (Polynomial discrimination $\implies$ Glivenko Cantelli)
Any function class with polynomial discrimination is Glivenko-Cantelli, ie $\|P_n f - Pf\|_{\mathcal{F}} = o_p(1)$.

Example 57 (Polynomial discrimination of indicators of half-intervals)

$$\mathcal{F} = \{\mathbf{1}[x < \theta] : \theta \in \mathbb{R}\}.$$

Then

$$|\mathcal{F}(x_1, \dots, x_n)| \leqslant n + 1$$

Which means that we have polynomial complexity of order $\nu = 1$. As a result:

$$P(\|\hat{F}_n - F\|_\infty \geqslant 8\sqrt{\frac{\log(n+1)}{n}}\delta) \leqslant \exp(-n\delta^2/2) \quad \forall \delta \geqslant 0,$$

and $\|\mathcal{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$ with a Borel-Cantelli.

10 VC Dimension

In the Glivenko-Cantelli case, we can directly calculate the Polynomial Discrimination. But it's not always so easy– VC dimension provides another method for *binary functions*.