STATS 300B Notes

Kenneth Tay

1 Preliminaries

- (Lec 1) **SLLN and CLT:** If $X \overset{iid}{\sim} P$, $\text{Cov}(X_i) = \Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)^T]$, $\mu = \mathbb{E}X_i$, then $\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i \overset{a.s.}{\to} \mu$ and $\dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^n (X_i - \mu) \overset{a.s.}{\to} \mathcal{N}(0, \Sigma)$.
- (Lec 1) **Continuous mapping theorem:** Let g be continuous on a set B such that $\mathbb{P}(X \in B) = 1$. Then $X_n \overset{a.s.}{\to} X \Rightarrow g(X_n) \overset{a.s.}{\to} g(X)$, $X_n \overset{P}{\to} X \Rightarrow g(X_n) \overset{P}{\to} g(X)$, and $X_n \overset{d}{\to} X \Rightarrow g(X_n) \overset{d}{\to} g(X)$.
- (Lec 1) **Slutsky's theorem:**
  - If c is constant, then $X_n \overset{d}{\to} c \iff X_n \overset{P}{\to} c$.
  - If $X_n \overset{d}{\to} X$ and $d(X_n, Y_n) \overset{P}{\to} 0$, then $Y_n \overset{d}{\to} X$.
  - If $X_n \overset{d}{\to} X$, $Y_n \overset{P}{\to} c$, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \overset{d}{\to} \begin{pmatrix} X \\ c \end{pmatrix}$.
  - If $X_n \overset{d}{\to} X$ and $Y_n \overset{d}{\to} c$, then $X_n + Y_n \overset{d}{\to} X + c$, $Y_n X_n \overset{d}{\to} cX$, and $\frac{X_n}{Y_n} \overset{d}{\to} \frac{X}{c}$ if $c \neq 0$. (This holds even for matrices.)
- (Lec 2) **Uniform tightness:** A collection of random vectors $\{X_\alpha\}_{\alpha \in A}$ is **uniformly tight** if

$$\limsup_{M \to \infty} \left[\sup_\alpha \mathbb{P}(\|X_\alpha\| \geq M) \right] = 0.$$

  A single random vector is uniformly tight. If $X_n \overset{d}{\to} X$, then $\{X_n\}$ is uniformly tight.
- (Lec 1) **O notation:** Let X_n be random vectors, R_n be random variables.
  - We say that $X_n = o_p(R_n)$ if there are random vectors Y_n such that $X_n = Y_n R_n$ and $Y_n \overset{P}{\to} 0$.
  - We say that $X_n = O_p(R_n)$ if there are random vectors Y_n such that $X_n = Y_n R_n$ and $Y_n = O_p(1)$, i.e. $\{Y_n\}$ uniformly tight.
  - $o_p(1) + o_p(1) = o_p(1)$.
  - $O_p(1) + o_p(1) = O_p(1)$.
  - $O_p(1) + O_p(1) = O_p(1)$.
  - $O_p(1) o_p(1) = o_P(1)$.
  - $[1 + o_p(1)]^{-1} = O_p(1)$.
  - $o_p(O_p(1)) = o_p(1)$.

- Let $R : \mathbb{R}^d \mapsto \mathbb{R}^k$ be a function with $R(0) = 0$, and assume $X_n \xrightarrow{P} 0$. If $R(h) = o(\|h\|^p)$ as $h \to 0$, then $R(X_n) = o_p(\|X_n\|^p)$. If $R(h) = O(\|h\|^p)$ as $h \to 0$, then $R(X_n) = O_p(\|X_n\|^p)$.

- (Lec 2) **Prohorov's theorem:** A collection of random vectors $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight iff it is sequentially compact for weak convergence, i.e. $\forall$ sequences $\{X_n\}_{n\in\mathbb{N}} \subset \{X_\alpha\}_{\alpha\in A}$, there exist a subsequence n_k and a random vector X such that $X_{n_k} \xrightarrow{d} X$.

- (Lec 2) **Portmanteau theorem:** Let X_n, X be random vectors. The following are equivalent:

  1. $X_n \xrightarrow{d} X$.
  2. $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all bounded and continuous f.
  3. $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for 1-Lipschitz f with $f \in [0, 1]$.
  4. $\liminf_{n\to\infty} \mathbb{E}(f(X_n)) \geq \mathbb{E}(f(X))$ for non-negative and continuous f.
  5. $\liminf_{n\to\infty} \mathbb{P}(X_n \in O) \geq \mathbb{P}(X \in O)$ for all open sets O.
  6. $\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \leq \mathbb{P}(X \in C)$ for all closed sets C.
  7. $\lim_{n\to\infty} \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$ for all sets B such that $\mathbb{P}(X \in \partial B) = 0$.

2 Delta Method

- (Lec 2) Let $r_n \to \infty$ be a deterministic sequence and $\phi : \mathbb{R}^d \mapsto \mathbb{R}^k$ be differentiable at θ. Assume that $r_n(T_n - \theta)$ converges in distribution to some random vector $T \in \mathbb{R}^d$. Then,

  1. $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$, and
  2. $r_n(\phi(T_n) - \phi(\theta)) - r_n\phi'(\theta)(T_n - \theta) \xrightarrow{P} 0$.

  Here $\phi'(\theta) \in \mathbb{R}^{k\times d}$ is the Jacobian matrix of derivatives $[\phi'(\theta)]_{ij} = \dfrac{\partial \phi_i(\theta)}{\partial \theta_j}$.

- (Lec 2) If $\phi'(\theta) = 0$, we can do a higher order Taylor expansion to get more power results/faster rate of convergence.

- (Lec 2) Let $r_n \to \infty$ be a deterministic sequence and $\phi : \mathbb{R}^d \mapsto \mathbb{R}^k$ be twice differentiable at θ such that $\nabla\phi(\theta) = 0$. Then,
$$r_n^2(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \frac{1}{2} T^T \nabla^2\phi(\theta) T,$$
  where $\nabla^2\phi(\theta)$ is the Hessian matrix.

- (Lec 5) **Restatement of CLT:** Assume $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$. Let $f : \mathcal{X} \mapsto \mathbb{R}^d$ with $P_{\theta_0} \|f\|_2^2 < \infty$. Then CLT says that under P_{θ_0}, $\sqrt{n}(P_n f - P_{\theta_0} f) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}_{\theta_0}(f))$.

- (Lec 5) **Delta method for method of moments:** Suppose that $e(\theta) = P_\theta f$ is one-to-one on an open set $\Theta \subseteq \mathbb{R}^d$ and continuously differentiable at θ_0 with non-singular derivative e'_{θ_0}. Assume also that $P_{\theta_0} \|f\|_2^2 < \infty$. Then $P_n f \in \mathrm{dom}\ (e^{-1})$ eventually, and if $\hat{\theta}_n = e^{-1}(P_n f)$, under P_{θ_0} we have
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, [e'(\theta_0)]^{-1}\mathrm{Cov}_{\theta_0}(f)([e'(\theta_0)]^{-1})^T\right).$$

3 Asymptotic Normality

Set-up: Model family $\{P_\theta\}_{\theta\in\Theta}$, where $\Theta \subseteq \mathbb{R}^d$. Assume P_θ has density p_θ w.r.t. some base measure μ. We denote the log-likelihood by $\ell_\theta(x) = \log p_\theta(x)$.

Say we observe $X_i \overset{iid}{\sim} P_{\theta_0}$, where θ_0 is unknown and we wish to estimate it.

- (Lec 3) **Score function:** $\nabla \ell_\theta(x) := \left[\dfrac{\partial}{\partial\theta_j}\log p_\theta(x)\right]_{j=1}^d \in \mathbb{R}^d$. Also written as $\dot{\ell}_\theta$. **We always have** $\mathbb{E}_\theta[\nabla\ell_\theta(x)] = 0$.
- (Lec 3) **Fisher information:** $I_\theta = \mathrm{Cov}_\theta \nabla\ell_\theta = \mathbb{E}_\theta\left[\nabla\ell_\theta\nabla\ell_\theta^T\right] = -\mathbb{E}_\theta[\nabla^2\ell_\theta]$.
  - (Lec 4) For a function h, $I(h(\theta)) = \dfrac{I(\theta)}{h'(\theta)^2}$.
  - (Lec 4) Fisher information is additive (if stuff is independent).
- (Lec 3) **MLE estimation:** The MLE estimate is
$$\hat{\theta}_n \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, P_n\ell_\theta(x) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i)$$
.
- **Functional invariance of MLE:** If $\hat\theta$ is MLE for θ, then for any function f, $f(\hat\theta)$ is MLE for $f(\theta)$.
- (Lec 3) **Consistency:** An estimator $\hat\theta_n$ is consistent if $\hat\theta_n \xrightarrow{P} \theta_0$ as $n\to\infty$.
- (Lec 3) **Identifiability:** A model $\{P_\theta\}_{\theta\in\Theta}$ is identifiable if $P_{\theta_1} \neq P_{\theta_2}$ for all $\theta_1, \theta_2 \in \Theta$ with $\theta_1 \neq \theta_2$.
- (Lec 3) **Consistency of MLE for finite Θ:** Suppose $\{P_\theta\}_{\theta\in\Theta}$ is identifiable and $|\Theta| < \infty$. Then, if $\hat\theta_n := \operatorname{argmax}_{\theta\in\Theta} P_n\ell_\theta(x)$, $\hat\theta_n \xrightarrow{P} \theta_0$ when $X_i \overset{iid}{\sim} P_{\theta_0}$.
- (Lec 4) **Asymptotic normality of the MLE:** Let $X_i \overset{iid}{\sim} P_{\theta_0}$, where $\theta_0 \in \operatorname{int}\Theta$. Assume that:
  1. $\ell_\theta(x) = \log p_\theta(x)$ is smooth enough that $\mathbb{E}_{\theta_0}[\nabla\ell_{\theta_0}\nabla\ell_{\theta_0}^T]$ exists,
  2. The Hessian $\nabla^2\ell_\theta(x)$ is $M(x)$-Lipschitz in θ, i.e. $\left\|\nabla^2\ell_{\theta_1}(x) - \nabla^2\ell_{\theta_2}(x)\right\|_{\mathrm{op}} \le M\left\|\theta_1-\theta_2\right\|$, with $\mathbb{E}_{\theta_0}[M(X)^2] < \infty$,
  3. The MLE $\hat\theta_n$ is consistent, i.e. $\hat\theta_n \xrightarrow{P} \theta_0$ under P_{θ_0}.

  Then, $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, I_{\theta_0}^{-1}\right)$, where $I_\theta = \mathbb{E}_\theta[\nabla\ell_\theta\nabla\ell_\theta^T]$ is the Fisher information.
  (Exponential family version of this theorem is in VdV Thm 4.6 p39.)
- (Lec 4) **Covariance lower bound:** For any decision procedure $\delta : \mathcal{X} \mapsto \mathbb{R}$ and any function $\psi : \mathcal{X} \mapsto \mathbb{R}$ ($\mathcal{X}$ is where the data lives, $\mathbb{R}$ is where the parameter lives), we have $\mathrm{Var}\ (\delta) \ge \dfrac{\mathrm{Cov}(\delta,\psi)^2}{\mathrm{Var}\ (\psi)}$. (Proof by Cauchy-Schwarz.)
- (Lec 4) **1-Dimensional information inequality:** Assume that $\mathbb{E}_\theta[\delta] = g(\theta)$ is differentiable at θ and density P_θ is regular enough so that we can interchange differentiation and integration. Then $\mathrm{Var}_\theta(\delta) \ge \dfrac{[g'(\theta)]^2}{I(\theta)}$.

  Implication: In 1 dimension, any unbiased estimator has MSE $\ge 1/I(\theta)$.

- (Lec 4) **Multi-dimensional covariance lower bound:** Assume we have $\delta : \mathcal{X} \mapsto \mathbb{R}^d$ and any function $\psi : \mathcal{X} \mapsto \mathbb{R}^d$ with $\mathbb{E}_\theta[\psi] = 0$. Let $\gamma = [\mathrm{Cov}(\delta, \psi_j)]_{j=1}^d$, and $C = \mathrm{Cov}_\theta(\psi) = \mathbb{E}_\theta[\psi\psi^T]$. Then $\mathrm{Var}_\theta(\delta) \geq \gamma^T C^{-1} \gamma$.
- (Lec 4) **Multi-dimensional information inequality:** Assume that $\mathbb{E}_\theta[\delta] = g(\theta) \in \mathbb{R}^d$, and that we have enough regularity so that we can interchange differentiation and integration. Then $\mathrm{Var}_\theta(\delta) \succeq \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$.

4 Efficiency of Estimators

- (Lec 5) An estimator $\hat{\theta}_n$ is **efficient** if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I_\theta^{-1})$ under P_θ.
- (Lec 5) **Asymptotic relative efficiency (ARE):** Let $\hat{\theta}_n$ and T_n be estimators of parameter $\theta \in \mathbb{R}$. Assume that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$. Let $m(n) \to \infty$ be such that $\sqrt{n}(T_{m(n)} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$. Then the **asymptotic relative efficiency** of $\hat{\theta}_n$ with respect to T_n is $\liminf_{n\to\infty} \dfrac{m(n)}{n}$.
- (Lec 5) Bigger ARE means that $\hat{\theta}_n$ is a better (more efficient) estimator than T_n.
- (Lec 5) ARE is related to the relative length of confidence intervals.
- (Lec 5) Suppose $\hat{\theta}_n$ and T_n are estimators of θ such that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ and $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \tau^2(\theta))$. Then the ARE of $\hat{\theta}_n$ with respect to T_n is $\dfrac{\tau^2(\theta)}{\sigma^2(\theta)}$. (In higher dimensions, it is roughly $\mathrm{tr}\left(\tau^2(\theta)(\sigma^2(\theta)^{-1}\right)$.)
- (Lec 6) **Super-efficiency:** An estimator $\hat{\theta}_n$ is super-efficient if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ under P_θ, with $\sigma^2(\theta) \leq I(\theta)^{-1}$ for all θ and $\sigma^2(\theta_0) < I(\theta_0)^{-1}$ for some θ_0.

5 U-Statistics (VdV Ch 12)

- (Lec 6) Let $h : X^r \mapsto \mathbb{R}$ be symmetric. For $X_i \overset{iid}{\sim} P$, define $\theta(P) := \mathbb{E}_P\left[h\left(X_1, \ldots, X_r\right)\right]$ and associated **U-statistic** $U_n := \dfrac{1}{\binom{n}{r}} \displaystyle\sum_{|\beta|=r, \beta\subseteq[n]} h(X_\beta)$, where β ranges over size r subsets of $[n] = \{1, ..., n\}$, $X_\beta = (X_{i_1}, ..., X_{i_r})$ for $\beta = (i_1, ..., i_r)$.
- (Lec 6) $\mathbb{E}_P[U_n] = \theta(P)$, i.e. the U-statistic is unbiased.
- (Lec 6) Let $h : X^r \mapsto \mathbb{R}$ be symmetric. For $c < r$, Define the following quantities:

$$h_c(X_1, \ldots, X_c) := \mathbb{E}\left[h\left(\underbrace{X_1, \ldots, X_c}_{\text{fixed}}, \underbrace{X_{c+1}, \ldots, X_r}_{\text{i.i.d. P}}\right)\right],$$
$$\hat{h}_c := h_c - \mathbb{E}[h_c] = h_c - \theta(P),$$
$$\zeta_c := \mathrm{Var}[h_c(X_1, \ldots, X_c)] = \mathbb{E}\left[\hat{h}_c^2\right].$$

- (Lec 7) If $\alpha, \beta \subseteq [n]$, $S = \alpha \cap \beta$, $c = |S|$, then $\mathbb{E}\left[\hat{h}(X_\alpha)\hat{h}(X_\beta)\right] = \zeta_c$.

- (Lec 7) Let U_n be an r^{th} order U-statistic. Then $\text{Var } U_n = \frac{r^2}{n}\zeta_1 + O(n^{-2})$.

- (Lec 7) **Projections:** Let $\mathcal{V}$ be a Hilbert space, and let $C \subseteq \mathcal{V}$ be a convex and closed set. Define the **projection of** w **onto** C as $\pi_C(w) := \underset{v \in C}{\operatorname{argmin}}\{\|w - v\|_2^2\}$.

  $\pi_C(w)$ exists, is unique, and is characterized by the inequality $\langle w - \pi_C(w), v - \pi_C(w)\rangle \leq 0$.

- (Lec 7) Suppose C is a linear subspace of $\mathcal{V}$. Then $\pi_C(w)$ is the projection of w onto C iff for all $v \in C$, $\langle w - \pi_C(w), v\rangle = 0$.

- (Lec 7, VdV Thm 11.1) If $\mathcal{S}$ is a linear subspace of $L_2(P)$, then $\hat{S} \in \mathcal{S}$ is the projection of $T \in L_2(P)$ onto $\mathcal{S}$ iff for all $S \in \mathcal{S}$, $\mathbb{E}[(T - \hat{S})S] = 0$.

  Every two projections of T onto $\mathcal{S}$ are a.s. equal. If the linear space $\mathcal{S}$ contains the constant variables, then $\mathbb{E}T = \mathbb{E}\hat{S}$ and $\text{Cov}(T - \hat{S}, S) = 0$ for every $S \in \mathcal{S}$.

- (Lec 7) Let T_n be statistics, and let $\hat{S}_n$ be the projections of T_n onto subspaces $\mathcal{S}_n$ which contain constant random variables.

$$\text{If } \frac{\text{Var } T_n}{\text{Var } \hat{S}_n} \to 1, \text{ then } \frac{T_n - \mathbb{E}T_n}{\sqrt{\text{Var } T_n}} - \frac{\hat{S}_n - \mathbb{E}\hat{S}_n}{\sqrt{\text{Var } \hat{S}_n}} \xrightarrow{P} 0.$$

- (Lec 7) **Hájek Projections**: Let $X_1, \ldots, X_n$ be independent. Let $\mathcal{S} = \left\{\sum_{i=1}^n g_i(X_i) : g_i \in L_2(P)\right\}$. If $\mathbb{E}T^2 < \infty$, then the projection $\hat{S}$ of T onto $\mathcal{S}$ is given by $\hat{S} = \sum_{i=1}^n \mathbb{E}[T \mid X_i] - (n-1)\mathbb{E}T$.

- (Lec 8) **Asymptotic normality of U-statistics:** Let h be a symmetric kernel (function) of order r with $\mathbb{E}[h^2] < \infty$, and let U_n be the associated U-statistic. Then $\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, r^2\zeta_1)$, where $\theta = \mathbb{E}[U_n] = \mathbb{E}h(X_1, \ldots, X_n)$.

6 Testing and Confidence Intervals

- (Lec 9) Let T_n be a sequence of tests for some model $\{P_\theta\}_{\theta \in \Theta}$, and let $H_0 : \theta \in \Theta_0 \subset \Theta$. Then T_n is **asymptotically level** α if

$$\lim_{n \to \infty} \sup_{\theta \in \Theta_0} P_\theta(T_n \text{ rejects } H_0) \leq \alpha.$$

- (Lec 8) Suppose $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1})$. Assume that I_θ is continuous and invertible. Let $C_{n,\gamma} := \left\{\theta : \mathbb{R}^d : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq \frac{\gamma}{n}\right\}$. Then as $n \to \infty$, $P_{\theta_0}\{\theta_0 \in C_{n,\gamma}\} \to \alpha$. $C_{n,\gamma}$ is called a **Wald confidence ellipsoid**. (We have $n(\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \xrightarrow{d} \chi_d^2$.)

- In one-dimension, the **Wald confidence interval** is $\hat{\theta}_n \pm \frac{1}{\sqrt{n}}\sqrt{I(\hat{\theta}_n)^{-1}} \cdot z^{1-\alpha/2}$.

- (Lec 9) **Wald tests:** Let $u_{d,\alpha}^2$ be the value such that $\mathbb{P}(\chi_d^2 \geq u_{d,\alpha}^2) = \alpha$. Then the Wald test rejects if $\hat{\theta}_n \notin C_{n,\gamma}$ with $\gamma = u_{d,\alpha}^2$.

  If H_0 is a point null (just θ_0), we can replace $I_{\hat{\theta}_n}$ with I_{θ_0} in the confidence ellipsoid.

- (Lec 8) **Generalized Likelihood Ratio Test:** Suppose we are testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta$, where Θ_0 is a strict subset of Θ. Define statistic

$$T(x) = \log \frac{\sup_{\theta \in \Theta} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)} = \log \frac{p(x, \hat{\theta}_{\text{MLE}})}{\sup_{\theta \in \Theta_0} p(x, \theta)}.$$

The Generalized Likelihood Ratio test rejects H_0 if $T(X) > t$.

- (Lec 8) **Wilk's Theorem:** Setting as above, where $\Theta_0 = \{\theta_0\}$ is a point null and $\Theta = \mathbb{R}^d$. Assume that the usual asymptotic normality conditions hold (log-likelihood is twice-differentiable, Hessian is Lipschitz-continuous). Let $X = (X_1, \ldots, X_n)$, where $X_i \overset{iid}{\sim} P_{\theta_0}$. Then $2T_n(X) \xrightarrow{d} \chi^2_d$ under the null.

- (Lec 9) Assume a point null $H_0 : \theta = \theta_0$. Under the CLT, $\sqrt{n} P_n \nabla \ell_{\theta_0} \xrightarrow{d} \mathcal{N}(0, I_{\theta_0})$, so $n(P_n \nabla l_{\theta_0})^T I_{\theta_0}^{-1} (P_n \nabla l_{\theta_0}) \xrightarrow{d} \chi^2_d$. The **Rao score test** for H_0 vs. $H_1 : \theta \neq \theta_0$ is reject if $(P_n \nabla l_{\theta_0})^T I_{\theta_0}^{-1} (P_n \nabla l_{\theta_0}) \geq u^2_{d,\alpha}/n$.

  Rao score test is useful if MLE is difficult to compute.

  ($\sqrt{n} P_n \nabla \ell_{\theta_0}$ could potentially be used for a one-sided test.)

6.1 Testing with Nuisance Parameters

Assume that our model family is parametrized by $\theta \in \mathbb{R}^p$. Let $\theta = (\eta, \nu)$, where $\eta \in \mathbb{R}^k$. Say we are only interested in η and not ν. We have to modify our tests above to account for the nuisance parameters.

Assume that we are testing $\eta = \eta_0$ vs. $\eta \neq \eta_0$.

- Write the Fisher information matrix in block form: $I_\theta = \begin{pmatrix} I_{\eta\eta} & I_{\eta\nu} \\ I_{\nu\eta} & I_{\nu\nu} \end{pmatrix}$. Then the upper left block of I_θ^{-1} is $(I_{\eta\eta} - I_{\eta\nu} I_{\nu\nu}^{-1} I_{\nu\eta})^{-1}$.

- **Wald test:** We now have $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} \mathcal{N}(0, (I_{\eta\eta} - I_{\eta\nu} I_{\nu\nu}^{-1} I_{\nu\eta})^{-1})$.

- **Score test:** Let $\hat{\nu}_0$ be the MLE for ν when η is fixed at η_0. Let $\nabla_\eta \ell(\eta, \nu)$ denote the gradient of the score function w.r.t. just the parameters we care about. Then the new score statistic is $Z_n = \sqrt{n} P_n \nabla_\eta \ell(\eta_0, \hat{\nu}_0)$. Under the null, $Z_n \xrightarrow{d} \mathcal{N}(0, I_{\eta\eta} - I_{\eta\nu} I_{\nu\nu}^{-1} I_{\nu\eta})$.

- **Likelihood ratio test:** $R_n = 2\ell(\hat{\theta}) - 2\ell(\eta_0, \hat{\nu}_0) \xrightarrow{d} \chi^2_{p-k}$.

  For more details, see TSH Thm 12.4.2 p515.

7 Uniform Laws of Large Numbers (ULLN)

- (TSH Thm 11.2.17 p441, HW1) **Glivenko-Cantelli Theorem:** Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with cdf F. Let $\hat{F}_n$ be the ECDF defined by $\hat{F}(t) = \frac{1}{n} \sum_{i=1}^n 1\{X_i \leq t\}$.

  Then $\sup_t |\hat{F}_n(t) - F(t)| \xrightarrow{a.s.} 0$.

- (Lec 10) X is a mean zero σ^2−**sub-Gaussian random variable** if $\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

- $\mathcal{N}(0, \sigma^2)$ is σ^2-sub-Gaussian.
- If $X \in [a, b]$, then X is $\dfrac{(b-a)^2}{4}$-sub-Gaussian.
- If $|X| \leq c$, then X is c^2-sub-Gaussian.
- If X_i's are independent σ_i^2-sub-Gaussian random variables, then $\sum X_i$ is a $\sum \sigma_i^2$-sub-Gaussian random variable.

- X is a **sub-Gaussian random vector with variance proxy** σ^2 if $\mathbb{E}X = 0$ and $u^T X$ is σ^2-sub-Gaussian for all unit vectors u on the unit sphere in $\mathbb{R}^d$.

- (Lec 11, HW1) If $X_1, \ldots, X_n$ are mean 0 σ^2-sub-Gaussian random variables, then $\mathbb{E}\left[\max_{1 \leq k \leq n} X_k\right] \leq \sqrt{2\sigma^2 \log n}$ and $\mathbb{E}\left[\max_{1 \leq k \leq n} |X_k|\right] \leq \sqrt{2\sigma^2 \log(2n)}$.

- (Lec 10) If X is σ^2-sub-Gaussian, then $\max[\mathbb{P}(X - \mathbb{E}X \geq t), \mathbb{P}(X - \mathbb{E}X \leq -t)] \leq \exp\left(-\dfrac{t^2}{2\sigma^2}\right)$.

- (Lec 10) **Hoeffding's inequality:** Let X_i's be independent σ_i^2-sub-Gaussian random variables. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left[\frac{-n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\right] \quad \text{for } t \geq 0,$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right) \leq \exp\left[\frac{-n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\right] \quad \text{for } t \leq 0.$$

  When we have $X_i \in [a, b]$ for all i, the bound on the RHS becomes $\exp\left[-\dfrac{2nt^2}{(b-a)^2}\right]$.

- (Lec 10) Setting for coverings and packings: Let (Θ, d) be a metric space with distance measure $d : \Theta \times \Theta \mapsto \mathbb{R}$.

- (Lec 10) For any $\varepsilon > 0$, $\{\theta_i\}_{i=1}^N$ is an ε-**cover** of Θ if $\min_i d(\theta, \theta_i) < \varepsilon$ for all $\theta \in \Theta$. (We do not require that $\theta_i \in \Theta$.)

- (Lec 10) For $\varepsilon > 0$, the **covering number** of Θ for metric d is $N(\Theta, d, \varepsilon) := \inf\{N : \exists \text{ an } \varepsilon-\text{cover } \{\theta_i\}_{i=1}^N \text{ of } \Theta\}$. $\log N(\Theta, d, \varepsilon)$ is called the **metric entropy**.

- (Lec 10) For any $\varepsilon > 0$, $\{\theta_i\}_{i=1}^M$ is an ε-**packing** of Θ if $\min_{i,j} d(\theta_i, \theta_j) > \varepsilon$. (We require $\theta_i \in \Theta$.)

- (Lec 10) For $\varepsilon > 0$, the **packing number** of Θ for metric d is $M(\Theta, d, \varepsilon) := \sup\{M : \exists \text{ an } \varepsilon-\text{packing } \{\theta_i\}_{i=1}^M \text{ of } \Theta\}$. $\log M(\Theta, d, \varepsilon)$ is called the **packing entropy**.

- (Lec 10, Quals Ex 4) For all ε, $M(2\varepsilon) \leq N(\varepsilon) \leq M(\varepsilon)$.

- (Lec 10) Let $\Theta \subseteq \mathbb{R}^d$ be compact. Then $N(\Theta, \|\cdot\|, \varepsilon) < \infty$ for any $\varepsilon > 0$.

- (Lec 10) **Balls in $\mathbb{R}^d$ with Euclidean norm:** If the ball has radius r, then $M(\Theta, \|\cdot\|, \varepsilon) \leq \left(1 + \dfrac{2r}{\varepsilon}\right)^d$, and $\left(\dfrac{r}{\varepsilon}\right)^d \leq N(\Theta, \|\cdot\|, \varepsilon) \leq \left(1 + \dfrac{2r}{\varepsilon}\right)^d$. Thus, $\log N(\Theta, \|\cdot\|, \varepsilon) \approx d \log\left(1 + \dfrac{r}{\varepsilon}\right)$.

- (Lec 10) Let $\mathcal{F} \subseteq \{f : \mathcal{X} \mapsto \mathbb{R}\}$ be a collection of functions with measure μ on $\mathcal{X}$. A set $\{[l_i, u_i]\}_{i=1}^N$ of functions $\mu_i, l_i : \mathcal{X} \to \mathbb{R}$ is an ε-**bracketing** of $\mathcal{F}$ in the $L_p(\mu)$ norm if

1. $\int [u_i(x) - l_i(x)]^p \, d\mu(x) \leq \varepsilon^p$ for all i, and
2. For all $f \in \mathcal{F}$, there is an i s.t. $l_i(x) \leq f(x) \leq u_i(x)$ for all x.

- (Lec 10) The **bracketing number** of $\mathcal{F}$ is $N_{[]}(\mathcal{F}, L_p(\mu), \varepsilon) := \inf \left\{ N : \exists \text{ an } \varepsilon\text{-bracketing of } \mathcal{F} \ \{[l_i, u_i]\}_{i=1}^N \right\}$.

- (HW7) The collection of functions $f : [0,1] \mapsto [0,1]$ which are 1-Lipschitz has bracketing number (w.r.t. sup norm) bounded by $\exp(C/\varepsilon)$.

- (Lec 10) Let $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$, where the m_θ are L-Lipschitz in θ, with $\mathbb{E}[L(X)^p] < \infty$. Then $N_{[]}(\mathcal{F}, L_p, \varepsilon \|L(X)\|_p) \leq N(\Theta, \|\cdot\|, \varepsilon/2)$.

- (Lec 10) **Uniform convergence with bracketing numbers:** Let $\mathcal{F}$ satisfy $N_{[]}(\mathcal{F}, L_1(P), \varepsilon) < \infty$. Then under i.i.d. sampling, $\sup\limits_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{P} 0$.

- (Lec 11) For a function class $\mathcal{F}$, we define the $\mathcal{F}$**-norm** $\|P_n - P\|_{\mathcal{F}} := \sup\limits_{f \in \mathcal{F}} |P_n f - Pf|$. $\mathcal{F}$ satisfies a **uniform law of large numbers** if $\lim\limits_{n \to \infty} \|P_n - P\|_{\mathcal{F}} = 0$.

- (Lec 11) A function class $\mathcal{F}$ is a **Glivenko-Cantelli (GC) class** w.r.t. P if $\|P_n - P\|_{\mathcal{F}} \xrightarrow{P} 0$.

- (Lec 11) **Symmetrization:** If $X_1, ..., X_n$ are random vectors in a vector space equipped with a norm $\|\cdot\|$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables which are independent of the X_i's, then for $p \geq 1$,
$$\mathbb{E}\left[\left\| \sum_{i=1}^n X_i - \mathbb{E}[X_i] \right\|^p \right] \leq 2^p \mathbb{E}\left[\left\| \sum_{i=1}^n \varepsilon X_i \right\|^p \right]$$

- (Lec 11) **Symmetrization inequality:** If $\mathcal{F}$ is a function class, then by symmetrization,
$$\frac{1}{2}\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf| \right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$
The term on the right is known as the **Rademacher complexity** of $\mathcal{F}$, denoted $R_n(\mathcal{F})$. (See HW6 for Rademacher complexity for some collections of functions.)

- (HW6) **Ledoux-Talagrand Rademacher contraction inequality:** Let $\phi \circ \mathcal{F} = \{h : h(x) = \phi(f(x)), f \in \mathcal{F}\}$. If ϕ is an L-Lipschitz function, then $R_n(\phi \circ \mathcal{F}) \leq L R_n(\mathcal{F})$.

- (Lec 11) Let (T, d) be a metric space. We say $\{X_t\}_{t \in T}$ is a **sub-Gaussian process** if $\log \mathbb{E}\left[\exp\left(\lambda(X_s - X_t)\right)\right] \leq \frac{\lambda^2 d(s,t)^2}{2}$ for all $\lambda > 0, s, t \in T$.

- (Lec 11) Gaussian processes are sub-Gaussian processes.

- (Lec 11) Let T be a vector space with a norm $\|\cdot\|$, $X_i \in \mathcal{X}$ be random variables and loss function $\ell : T \times \mathcal{X} \mapsto \mathbb{R}$ be Lipschitz in its first argument, i.e. $|\ell(s,x) - \ell(t,x)| \leq ||t - s||$ for all $x \in \mathcal{X}, s, t \in T$. If we define $Z_t = \sum_{i=1}^n \epsilon_i \ell(t, x_i)$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables, then the stochastic process $\{X_t\}_{t \in T}$ is $n||\cdot||^2$-sub-Gaussian.

- (Lec 11) **Entropy integral:** For a metric space (T, d), the entropy integral is defined as $J(T, d) := \int_0^{\text{diam } T} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon$. ($N$ is covering number.)

- (Lec 11) **Dudley's Theorem:** If $\{X_t\}_{t\in T}$ is a separable sub-Gaussian process for metric $d(\cdot,\cdot)$, then $\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq CJ(T,d)$, for some numerical constant C. (C can be taken to be $4\sqrt{2}$.)

- (Lec 12) **Empirical norm:** For empirical distribution P_n, let $L_p(P_n)$ be the L_p norm w.r.t. P_n, i.e. $\|f\|_{L_p(P_n)} = \left[\frac{1}{n}\sum_{i=1}^n |f(X_i)|^p\right]^{1/p}$. We often use $L_2(P_n)$ for symmetrized processes.

- (Lec 12) If $\sqrt{n}P_n^o f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)$, then $f \mapsto \sqrt{n}P_n^o f$ is an $\|\cdot\|^2_{L_2(P_n)}$ sub-Gaussian process, so

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sqrt{n}P_n^o f\right| \Big| X_1\dots,X_n\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum \varepsilon_i f(X_i)\right| \Big| X_1\dots,X_n\right] \leq C\int_0^\infty \sqrt{\log N(\mathcal{F}, L_2(P_n),\varepsilon)}d\varepsilon.$$

- (Lec 12) **ULLNs with entropies:** For $M < \infty$, let $f_M(x) = f(x)1_{\{|f(x)|\leq M\}}$. For a collection of functions $\mathcal{F}$ with envelope F (i.e. $|f(x)|\leq F(x)$ for all x), let $\mathcal{F}_M := \{f_M : f\in\mathcal{F}\}$.

  If $\sqrt{\log N(\mathcal{F}_M, L_1(P_n),\varepsilon)} = o_p(n)$ for all $M<\infty$ and $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{P} 0$, ie. $\mathcal{F}$ is G.C. class.

- (VdV Thm 19.4) Every class $\mathcal{F}$ of measurable functions such that $N_{[]}(\mathcal{F}, L_1(P),\varepsilon) < \infty$ for every $\varepsilon > 0$ is a P-G.C. class.

- (VdV Thm 19.5) Every class $\mathcal{F}$ of measurable functions with $\int_0^1 \sqrt{\log N_{[]}(\mathcal{F}, L_2(P),\varepsilon)d\varepsilon} < \infty$ is P-Donsker.

8 Concentration Inequalities

- (Lec 10) If X is σ^2-sub-Gaussian, then $\max[\mathbb{P}(X-\mathbb{E}X\geq t), \mathbb{P}(X-\mathbb{E}X\leq -t)] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

- (Lec 10) **Hoeffding's inequality:** Let X_i's be independent σ_i^2-sub-Gaussian random variables. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left[\frac{-nt^2}{2\sum_{i=1}^n\sigma_i^2}\right] \quad \text{for } t\geq 0,$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq t\right) \leq \exp\left[\frac{-nt^2}{2\sum_{i=1}^n\sigma_i^2}\right] \quad \text{for } t\leq 0.$$

  When we have $X_i\in[a,b]$ for all i, the bound on the RHS becomes $\exp\left[-\frac{2nt^2}{(b-a)^2}\right]$.

- (TSH Thm 11.2.18 p442) **Dvoretzky-Kiefer-Wolfowitz (DKW) inequality:** Suppose $X_1,\dots,X_n$ are i.i.d. real-valued random variables with cdf F. Let $\hat{F}_n$ be the ECDF. Then, for any $d > 0$ and any positive n,

$$\mathbb{P}\left(\sup_t |\hat{F}_n(t) - F(t)| > d\right) \leq 2\exp(-2nd^2).$$

- (HW6) **Azuma's inequality:** A martingale $\{Z_k\}$ adapted to $\{X_1, \dots, X_k\}$ is σ_k^2**-sub-Gaussian** if for $\Delta_k = Z_k - Z_{k-1}$, we have $\mathbb{E}\left[\exp(\lambda\Delta_k) \mid \mathcal{F}_{k-1}\right] \le \exp\left(\frac{\lambda^2\sigma_k^2}{2}\right)$.

  Let Δ_k be a σ_k^2-sub-Gaussian martingale difference sequence with $Z_k = \sum_{i=1}^{k} \Delta_i$. Then Z_k is $\sum_{i=1}^k \sigma_i^2$-sub-Gaussian, and for $t \ge 0$,

$$\mathbb{P}(Z_k \ge t) \vee \mathbb{P}(Z_k \le -t) \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^k \sigma_i^2}\right).$$

- (HW6) **McDiarmid's Inequality:** Let $X_1, \dots, X_n$ be independent random variables. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a measurable function such that there exist $c_1, \dots, c_n$ with the property that for all $x_1, \dots, x_n$, $x_1', \dots, x_n'$, $|f(x_1, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)| \le c_i$.

  Let $W = f(X_1, \dots, X_n)$. Then for all $t > 0$,

$$P(W - \mathbb{E}W \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right), \qquad P(W - \mathbb{E}W \le -t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

- (HW6) Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \mapsto \mathbb{R}$, and let $R_n(\mathcal{F})$ be its Rademacher complexity. If $\mathcal{F}$ satisfies envelope condition $\sup_{x\in\mathcal{X}} \sup_{f\in\mathcal{F}} |f(x) - Pf| \le M$, then

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n f - Pf| \ge 2R_n(\mathcal{F}) + t\right) \le 2\exp\left(-\frac{cnt^2}{M^2}\right),$$

  for some numerical constant c and for all $t \ge 0$. (We can take $c = 1/2$.)

  Thus, if $R_n(\mathcal{F}) = o(1)$, then $\mathcal{F}$ is G.C. class.

- (Tail Bounds Thm 2.4) Let $(X_1, \dots, X_n)$ be i.i.d. standard Gaussian random variables, and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be L-Lipschitz w.r.t. Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is L^2-sub-Gaussian, and for all $t \ge 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \ge t\right) \le 2\exp\left(-\frac{t^2}{2L^2}\right).$$

9 VC Dimension

- (Lec 12) Let $\mathcal{C}$ be a collection of sets and $X = \{X_1, \dots, X_n\}$ be a collection of points. A vector $y \in \{+1, -1\}^n$ is a labeling of X. We say that $\mathcal{C}$ **shatters** X if for all labelings y of X, $\exists$ a set $A \in \mathcal{C}$, i.e., $X_i \in A$ if $y_i = 1$ and $X_i \notin A$ if $y_i = -1$.

- (Lec 12) The **VC dimension** of a collection of sets, $VC(\mathcal{C})$, is the size of the largest set $\{x_1, \dots, x_n\}$ s.t. $\mathcal{C}$ shatters $\{x_1, \dots, x_n\}$.

  The **subgraph** of a function: $\mathcal{X} \mapsto \mathbb{R}$ is sub $f := \{(x,t) : t < f(x)\} = (\text{epi } f)^c$ (the part of $\mathcal{X} \times \mathbb{R}$ below the graph of $f(x)$).

  $\mathcal{F}$ is a **VC-class/VC-subgraph-class** if sub $f : f \in \mathcal{F}$ is VC.

  - Half-spaces in $\mathbb{R}^d$ have $VC(\mathcal{C}) = d + 1$.
  - (Lec 13) Let $\mathcal{F} = \{f = \langle\theta, x\rangle : \theta \in \mathbb{R}^d\}$. Then $VC(\mathcal{F}) \le d + 2$.

- (Lec 13) If $\mathcal{F}$ is a linear space of functions with $\dim \mathcal{F} < \infty$, then $VC(\mathcal{F}) = O(\dim \mathcal{F})$.
- (Lec 13) If $\mathcal{C}$ and $\mathcal{D}$ are VC classes of sets, then $\mathcal{C} \cap \mathcal{D}$ and $\mathcal{C} \cup \mathcal{D}$ are as well.
- (Lec 13) If $\mathcal{F}$ is a VC class of functions and $\phi : \mathbb{R} \mapsto \mathbb{R}$ is monotone, then $\{\phi \circ f : f \in \mathcal{F}\}$ is VC.

- (Lec 12) Let $\Delta_n(\mathcal{C}, \{x_1, \ldots, x_n\}) :=$ the number of labellings $\mathcal{C}$ realizes on $\{x_i\}$. Then $VC(\mathcal{C}) := \sup\{n \in \mathbb{N} : \max\limits_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_1, \ldots, x_n\}) = 2^n\}$.

- (Lec 12) **Sauer-Shelah Lemma:** For any class $\mathcal{C}$, $\max\limits_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_i\}) \leq \sum\limits_{k=0}^{VC(\mathcal{C})} \binom{n}{k} = O(n^{VC(\mathcal{C})})$.

  Consequently, if $\sup\limits_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_i\}) < 2^n$, then $\Delta_n(\mathcal{C}, \{x_i\})$ is polynomial in n.

- (Lec 13) **VC bound on uniform covering number:** For sets A and B, define $\|A - B\|_{L_r(P)} = \|1_A - 1_B\|_{L_r(P)} = \left(\int |1_A - 1_B|^r dP\right)^{1/r}$. Then there is a universal constant $K < \infty$ s.t. for all $\varepsilon > 0$,

$$\sup_P N(\mathcal{C}, L_r(P), \epsilon) \leq K \cdot VC(\mathcal{C}) \cdot (4e)^{VC(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{r \cdot VC(\mathcal{C})},$$

  which implies that $\log N(\mathcal{C}, L_r(P), \epsilon) \leq c \cdot r \cdot VC(\mathcal{C}) \cdot \log(1/\epsilon)$.

- (Lec 13) **Using VC to get GC Theorem:** Let $\mathcal{F} = \{f(x) = 1_{x \leq t}, t \in \mathbb{R}^d\}$. Then $VC(\mathcal{F}) = O(d)$, implying $\sup\limits_P \log N(\mathcal{F}, L_2(P), \varepsilon) \leq K d \log(1/\varepsilon)$.

- (Lec 13) If $VC(\mathcal{F}) < \infty$ and $\mathcal{F}$ has envelope $F : \mathcal{X} \mapsto \mathbb{R}_+$, then

$$\sup_P N(\mathcal{F}, L_r(P), ||F||_{L_r(P)}\varepsilon) \leq \text{const} \cdot VC(\mathcal{F})(16e)^{VC(\mathcal{F})}(\frac{1}{\varepsilon})^{rVC(\mathcal{F})}.$$

10 Convergence in Distribution & Uniform CLTs

- (Lec 13) Let $\mathbb{D}$ be a metric space. X is a **random variable on** $\mathbb{D}$ if $X : \Omega \mapsto \mathbb{D}$.

- (Lec 13) Say X is $\mathbb{D}$-valued. Given sequence $X_n : \Omega_n \mapsto \mathbb{D}$, we say that $X_n \xrightarrow{d} X$ if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded and continuous $f : \mathbb{D} \mapsto \mathbb{R}$ (even Lipschitz).

- (Lec 13) Let (T, d) be a compact metric space. Let $L_\infty(T)$ denote the set of bounded functions $f : T \mapsto \mathbb{R}$. For $f, g \in L_\infty(T)$, define $\|f - g\|_\infty = \sup\limits_{t \in T} |f(t) - g(t)|$. Let $\ell : T \times \mathcal{X} \mapsto \mathbb{R}$ be continuous in t. Let $X, X_1, \ldots, X_n$ be $\mathcal{X}$-valued random variables. Define

$$Z_n(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\ell(\cdot, X_i) - \mathbb{E}\ell(\cdot, X)].$$

  Then Z_n is a $L_\infty(T)$-valued random variable. (Since $t \mapsto Z_n(t)$ is continuous, we have $\sup\limits_{t \in T} |Z_n(t)| < \infty$.)

  - If T_0 is a countable and dense subset of T, then Z_n is completely determined by $\{Z_n(t), t \in T_0\}$.
  - For fixed $t_1, \ldots, t_k$, by the CLT we get $\begin{pmatrix} Z_n(t_1) & \ldots Z_n(t_k) \end{pmatrix}^T \xrightarrow{d} \mathcal{N}\left(0, (\text{Cov}(\ell(t_i, X), \ell(t_j, X)))_{i,j=1}^k\right)$.

- (Lec 13) A random variable $X : \Omega \mapsto \mathbb{D}$ is **tight** if for all $\varepsilon > 0$, there is a compact set $K \subseteq \mathbb{D}$ such that $\mathbb{P}(X \notin K) < \varepsilon$.

  A sequence random variables $X_n : \Omega \mapsto \mathbb{D}$ is **asymptotically tight** if for all $\varepsilon > 0$, there is a compact set $K \subseteq \mathbb{D}$ such that $\limsup_{n\to\infty} \mathbb{P}(X_n \notin K^\delta) \leq \varepsilon$ for all $\delta > 0$. (Here, $K^\delta = \{x : d(x, K) < \delta\}$.)

  (Note: X_n individually tight does **not** imply that $\{X_n\}$ is asymptotically tight.)

- (Lec 13) **Prohorov's Theorem:** Let $X_n : \Omega \mapsto \mathbb{D}$ and $X : \Omega \mapsto \mathbb{D}$.

  1. If $X_n \xrightarrow{d} X$, where X is tight, then $\{X_n\}$ is asymptotically tight.
  2. If $\{X_n\}$ is asymptotically tight, then there is a subsequence $\{n_k\}$ and a tight $X : \Omega \mapsto \mathbb{D}$ such that $X_{n_k} \xrightarrow{d} X$.

- (Lec 13) For a function $f : T \mapsto \mathbb{R}$, its **modulus of continuity** is $w_f(\delta) := \sup_{d(s,t)<\delta} |f(t) - f(s)|$.

- (Lec 13) A collection of functions $\mathcal{F}$ is **uniformly equicontinuous** if $\lim_{s\downarrow 0} \sup_{f\in\mathcal{F}} w_f(\delta) = 0$.

- (Lec 13) **Arzelà-Ascoli Theorem:** Let (T, d) be a compact metric space. Let $\mathcal{C}(T, \mathbb{R})$ be the set of continuous functions $f : T \mapsto \mathbb{R}$. Then the following are equivalent:

  1. $\mathcal{F} \subseteq \mathcal{C}(T, \mathbb{R})$ is compact (or equivalently, sequentially compact).
  2. $\mathcal{F}$ is uniformly equicontinuous and there is a $t_0 \in T$ s.t. $\sup_{f\in\mathcal{F}} |f(t_0)| < \infty$.

- (Lec 14) Let $\{X_n\}$ be $L_\infty(T)$-valued random variables. (Recall $L_\infty(T)$ is the set of bounded functions $f : T \mapsto \mathbb{R}$.) We say that $\{X_n\}$ are **asymptotically equicontinuous** if for all $\eta, \varepsilon > 0$, there is a finite partition $T_1, \ldots, T_k$ of T such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_i \sup_{s,t\in T_i} |X_{n,s} - X_{n,t}| \geq \varepsilon\right) \leq \eta.$$

- (Lec 14) Let $Z_i \in \mathbb{R}^d$ with $Z_i \overset{iid}{\sim} P$. Assume that $\mathbb{E}[\|Z_i\|^2] < \infty$ and $\mathbb{E}[Z_i] = 0$.

  Let T be a compact subset of $\mathbb{R}^d$. For $t \in T$, define $X_{n,t} := \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^n Z_i^T t$. Then $\{X_n\}$ is asymptotically equicontinuous.

- (Lec 14) The following are equivalent:

  (i) $X_i \in L_\infty(T)$ and $X_n \xrightarrow{d} X \in L_\infty(T)$, where X is tight.

  (ii) (a) Finite dimensional convergence (FIDI): $(X_{n,t_1}, \ldots, X_{n,t_k}) \xrightarrow{d}$ something for any $t_1, \ldots, t_k \in T$, and $k < \infty$.

  (b) X_n are (asymptotically) stochastically equicontinuous.

- (Lec 15) **Uniform limits via entropy integral:** Suppose (T, d) is a totally bounded metric space with $\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{d(s,t)\leq\delta} |X_{n,s} - X_{n,t}| \geq \epsilon\right) = 0$, and we have FIDI (finite dimensional convergence of X_n to X). Then $X_n \xrightarrow{d} X$ in $L_\infty(T)$.

- (Lec 15) **Donsker class:** Consider a collection of functions $\mathcal{F}$. $L_\infty(\mathcal{F})$ is the collection of bounded functionals $m : \mathcal{F} \mapsto \mathbb{R}$. The process $\sqrt{n}(P_n - P)$ is a member of $L_\infty(\mathcal{F})$.

  $\mathcal{F}$ is P-**Donsker** if the empirical process $\mathbb{G}_n = \sqrt{n}(P_n - P)$ converges to a tight limit in $L_\infty(\mathcal{F})$.

- (Lec 15, VdV Thm 19.3) **Donsker's Theorem:** In the above setting, the limit $\mathbb{G}_n$, which we denote by $\mathbb{G}_P$ must be a Gaussian process because by the CLT, $\sqrt{n}(P_n - P)f \xrightarrow{d} \mathcal{N}(0, \text{Var}_P(f))$. We have

$$\mathbb{E}[\mathbb{G}_P f] = 0, \qquad \mathbb{E}[\mathbb{G}_P f \mathbb{G}_P g] = \text{Cov}_P(f, g) = P(fg) - Pf \cdot Pg.$$

- (Lec 15) **$\mathbb{P}$-Brownian bridge:** Let $F(t) = \mathbb{P}(X \leq t)$, $F_n(t) = \mathbb{P}_n(X \leq t)$ and $\mathcal{F} = \{1\{\cdot \leq t\}, t \in \mathbb{R}\}$. Then $\sqrt{n}(F_n(\cdot) - F(\cdot)) \xrightarrow{d} \mathbb{G}_P$ in $L_\infty(\mathbb{R})$. We have $\text{Cov}(1\{X \leq t\}, 1\{X \leq s\}) = F(s \wedge t) - F(s)F(t)$.

- (Lec 15) **Uniform CLT via entropy integral:** Let $\mathcal{F}$ be a collection of functions. Assume there exists an envelope function $B : X \mapsto \mathbb{R}_+$ such that $\mathbb{P}[B^2] < \infty$ and

$$\int_0^\infty \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|B\|_{L_2(Q)}\, \varepsilon)} d\varepsilon < \infty,$$

where the supremum is taken over all Q such that $Q[F^2] > 0$. Then $\mathcal{F}$ is $\mathbb{P}$-Donsker.

- (HW7) **Kolmogorov-Smirnov statistic:** Let $X_1, \ldots, X_m \overset{iid}{\sim} F$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} G$. The K-S statistic is the sup distance between the 2 empirical distributions, i.e. $K_{m,n} = \|F_m - G_n\|_\infty$. As $m, n \to \infty$, $\sqrt{\dfrac{mn}{m+n}} K_{m,n} \xrightarrow{d} \|\mathbb{G}\|_\infty$, where $\mathbb{G}$ is the Brownian bridge generated from F.

11 Modulus of Continuity

- (Lec 15) Say we have a criterion function $m_\theta : \mathcal{X} \mapsto \mathbb{R}$. Let $M_n(\theta) = \mathbb{P}_n m_\theta$ and $M(\theta) = \mathbb{P} m_\theta$. Then the **M-estimator** is $\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} M_n(\theta)$. (For example, maximum likelihood uses $m_\theta(x) = \log p_\theta(x)$.)

- (Lec 15) Say we have a criterion function $\psi_\theta : \mathcal{X} \mapsto \mathbb{R}$. Let $\Psi_n(\theta) = \mathbb{P}_n \psi_\theta$ and $\Psi(\theta) = \mathbb{P}\psi_\theta$. Then the **Z-estimator** is $\hat{\theta}_n$ such that $\Psi(\hat{\theta}_n) = 0$. (For example, maximum likelihood uses $\psi_\theta(x) = \nabla_\theta \log p_\theta(x)$.)

- (Lec 15, VdV Thm 5.7) **Consistency of M-estimators (Argmax consistency theorem):** Suppose we have $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$, and that for all $\varepsilon > 0$, $\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0)$.

  Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

  See VdV Thm 5.9 p46 for corresponding theorem for Z-estimators.

- (Lec 15) **Idea:** If $M(\theta)$ shrinks quickly away from θ_0 but $[M_n(\theta) - M(\theta)] - [M_n(\theta_0) - M(\theta_0)]$ is well-behaved (i.e. small error), then the M-estimator cannot be bad.

- (Lec 16) **Theorem:** Suppose $M(\theta_0) \geq M(\theta) + d(\theta, \theta_0)^2$ near θ_0. Let ϕ be such that $\phi(c\delta) \leq c^\alpha \phi(\delta)$ for some $\alpha \in (0, 2)$. Assume that we have a bound on the modulus of continuity:

$$\mathbb{E}\left[\sup_{d(\theta, \theta_0) \leq \delta} \Big|[M_n(\theta) - M(\theta)] - [M_n(\theta_0) - M(\theta_0)]\Big|\right] \leq \frac{\phi(\delta)}{\sqrt{n}}.$$

  Let $r_n \to +\infty$ such that $r_n^2 \phi\left(\dfrac{1}{r_n}\right) \leq \sqrt{n}$. If $\hat{\theta}_n \xrightarrow{P} \theta_0$, then $r_n d(\hat{\theta}_n, \theta_0) = O_P(1)$.

  - If $\phi(\delta) = \delta^\alpha$, we can solve to get $r_n = n^{\frac{1}{2(2-\alpha)}}$.
  - More generally, if we have $M(\theta_0) \geq M(\theta) + d(\theta, \theta_0)^\beta$, we could choose r_n so that $r_n^\beta \left(\dfrac{1}{r_n}\right) \leq \sqrt{n}$. We would then obtain $r_n = n^{\frac{1}{2(\beta-\alpha)}}$. (We need $\beta > \alpha$ in order for $r_n \to \infty$.)

12 Asymptotic Testing

- (Lec 17) A sequence of tests based on statistics T_n and rejection regions K_n is **asymptotically level (size)** α if $\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T_n \in K_n) \leq \alpha$.

- (Lec 17) Suppose that there exists a mean function $\mu(\theta)$ and a variance function $\sigma^2(\theta)$ such that $\sqrt{n}\left(\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)}\right) \xrightarrow[\theta_n]{d} \mathcal{N}(0,1)$. Suppose we are testing $\theta = 0$ vs. $\theta > 0$. If $\mu'(0)$ exists, $\sigma\left(\frac{h}{\sqrt{n}}\right) \to \sigma(0)$ as $n \to \infty$, then the level α test rejecting large values of $\sqrt{n}\frac{T_n - \mu(0)}{\sigma(0)}$ satisfies:

$$\text{Power } \pi_n\left(\frac{h}{\sqrt{n}}\right) \xrightarrow{n\to\infty} 1 - \Phi\left(z_\alpha - h\frac{\mu'(0)}{\sigma(0)}\right),$$

  where Φ is the standard normal distribution function.

- (Lec 17) **Slope of a test:** If $\sqrt{n}\left(\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)}\right) \xrightarrow[\theta_n]{d} \mathcal{N}(0,1)$ where $\theta_n = \frac{h}{\sqrt{n}}$, the slope of the tests T_n is defined as $\frac{\mu'(0)}{\sigma(0)}$.

  A bigger slope means a better test.

- (Lec 17) **Distinguishing numbers:** For $\nu \in \mathbb{N}$, consider a sequence of tests $H_0 : \theta = 0$ vs. $H_1 : \theta = \theta_\nu$, where $\theta_\nu \to 0$ as $\nu \to \infty$. Fix a level α and power $\beta \in (\alpha, 1)$. Define the **distinguishing number** $n_\nu := \inf\{n \in \mathbb{N} : \pi_n(0) \leq \alpha, \pi_n(\theta_\nu) \geq \beta\}$, i.e. the smallest number of observations necessary to distinguish H_0 from H_1 at level α and power β.

- (Lec 17) **ARE/Pitman Efficiency:** Let tests $T_n^{(1)}, T_n^{(2)}$ have distinguishing numbers $n_\nu^{(1)}, n_\nu^{(2)}$, respectively. Then the ARE/Pitman efficiency of $T^{(1)}$ relative to $T^{(2)}$ is defined as $\lim_{\nu\to\infty} \frac{n_\nu^{(2)}}{n_\nu^{(1)}}$.

  Larger ARE means $T^{(1)}$ is better than $T^{(2)}$.

- (Lec 17) Let models $\{P_{n,\theta}\}_{\theta\geq 0}$ satisfy $\lim_{\theta\to 0}\|P_{n,\theta} - P_{n,0}\|_{TV} = 0$ for every n. Let tests $T^{(1)}$, $T^{(2)}$ be such that as $\theta_n \downarrow 0$,

$$\sqrt{n}\left(\frac{T_n^{(i)} - \mu_i(\theta_n)}{\sigma_i(\theta_n)}\right) \xrightarrow[\theta_n]{d} \mathcal{N}(0,1),$$

  where $i \in \{1,2\}$, σ_i is continuous at 0, $\sigma_i(0) > 0$ and $\mu_i'(0) > 0$. Then the ARE of tests rejecting $H_0 : \theta = 0$ against $H_1 : \theta > 0$ when $T_n^{(i)}$ is large ($T_n^{(1)}$ relative to $T_n^{(2)}$) is

$$\left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2.$$

- See TSH E.g. 13.2.2 p537 for the ARE of t-test, Wilcoxon test and sign test in location model.

- (Lec 18) Let M be the joint measure (law) of the pair $(X, V) := \left(X, \frac{dQ}{dP}\right)$ under distribution P (so M is defined on $\mathcal{X} \times \mathbb{R}_+$). Then $V \geq 0$, $\mathbb{E}_M[V] = 1$, and $Q(B) = \mathbb{E}_P\left[1_{\{B\}}(X)\frac{dQ}{dP}\right] = \mathbb{E}_P[1_{\{B\}}(X)V] = \int_{B\times\mathbb{R}_+} V dM(x,v)$.

- (Lec 18) **Contiguity:** A sequence $\{Q_n\}$ of distributions is **contiguous** w.r.t. $\{P_n\}$, written $Q_n \triangleleft P_n$, if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for any sequence of sets A_n. Sequences $\{Q_n\}$ and $\{P_n\}$ are **mutually contiguous**, written $Q_n \triangleleft\triangleright P_n$, if $Q_n \triangleleft P_n$ and $P_n \triangleleft Q_n$.

- (HW9) If $\|P_n - Q_n\|_{TV} \to 0$, then P_n and Q_n are mutually contiguous.

- (Lec 18) Even if $Q_n \not\ll P_n$, we can still consider $Q_n = Q_n^{\parallel} + Q_n^{\perp}$, where $Q_n^{\parallel} \ll P$ and $Q_n^{\perp} \perp P$. If we define $\frac{dQ_n}{dP_n} := \frac{dQ_n^{\parallel}}{dP_n}$, then $\frac{dQ_n}{dP_n} \geq 0$ and $\mathbb{E}_{P_n}\left[\frac{dQ_n}{dP_n}\right] = Q_n^{\parallel}(\Omega) = 1 - Q_n^{\perp}(\Omega) \leq 1$. Thus, under P_n, the sequence $\frac{dQ_n}{dP_n}$ is tight.

- (Lec 18) We can always assume without loss of generality that P_n and Q_n all have densities p_n and q_n with respect to some finite base measure μ. One suitable measure is $\mu = \sum_{n=1}^{\infty} 2^{-n}(P_n + Q_n)$, which has total mass 2.

  We can also without loss of generality take $\dfrac{dQ_n}{dP_n} = \dfrac{q_n}{p_n}$, so that $\displaystyle\int \frac{q_n}{p_n} dP_n = \int q_n 1_{\{p_n > 0\}} d\mu = Q_n^{\parallel}(\Omega)$.

- We can think of $\dfrac{dP_n}{dQ_n}$ as a function $f_n : \mathcal{X} \mapsto \mathbb{R}_+$. When we say $\dfrac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} U$, what we mean is U is the limiting distribution of $f_n(X_n)$, where $X_n \sim Q_n$.

- (Lec 18, VdV Lem 6.4) **Le Cam's First Lemma:** The following are equivalent:
  1. $Q_n \triangleleft P_n$.
  2. If $\dfrac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} U$ along a subsequence, then $\mathbb{P}(U > 0) = 1$.
  3. If $\dfrac{dQ_n}{dP_n} \xrightarrow[P_n]{d} V$ along a subsequence, then $\mathbb{E}[V] = 1$.
  4. If $T_n \xrightarrow{P_n} 0$, then $T_n \xrightarrow{Q_n} 0$.

- (Lec 18, VdV Eg 6.5) **Asymptotic log normality**: Suppose that $\log \dfrac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} \mathcal{N}(\mu, \sigma^2)$. Then $Q_n \triangleleft P_n$. Further, $P_n \triangleleft Q_n$ iff $\mu = -\dfrac{\sigma^2}{2}$.

- (Lec 18) **Smooth likelihoods:** Suppose $\{P_\theta\}_{\theta \in \Theta}$ has densities p_θ which are smooth enough in θ such that $\log p_\theta$ has a Taylor expansion around $\theta_0 \in \operatorname{int} \Theta$. Then $P_{\theta_0}^n \triangleleft\triangleright P_{\theta_0 + h/\sqrt{n}}^n$.

- (Lec 18, VdV Thm 6.6) Let P_n, Q_n be distributions on $X_n \in \mathbb{R}^d$. If $Q_n \triangleleft P_n$ and $\left(X_n, \dfrac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} (X, V)$, then $L(B) := \mathbb{E}[1\{B\}(X)V]$ is a probability measure (ie. $\mathbb{E}[V] = 1$ and $V \geq 0$) and $X_n \xrightarrow[Q_n]{d} W$, where $W \sim L$.

- (Lec 18, TSH Cor 12.3.2 p500, VdV Eg 6.7) **Le Cam's Third Lemma:** If $\left(X_n, \log \dfrac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} \mathcal{N}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$, then $P_n \triangleleft\triangleright Q_n$ and $X_n \xrightarrow[Q_n]{d} \mathcal{N}(\mu + \tau, \Sigma)$.

- (Lec 19) **Hellinger distance:** For probability distributions P and Q with densities p and q w.r.t. some dominating μ, we define $d_{hel}^2(P, Q) = \dfrac{1}{2} \displaystyle\int (\sqrt{p} - \sqrt{q})^2 d\mu$.

  - $d^2_{hel}(P,Q) = 1 - \int \sqrt{pq}d\mu$.
  - (HW8) $d^2_{hel}(P,Q) \le \|P-Q\|_{TV} \le d_{hel}(P,Q)\sqrt{2-d^2_{hel}(P,Q)}$.
  - (Lec 19) $d^2_{hel}(P^n,Q^n) = 1 - \left(1-d^2_{hel}(P,Q)\right)^n$.
  -

- (Lec 19) When testing P_0 vs. P_1, the associated **best error** is $\inf_{\Phi:\mathcal{X}\mapsto\{0,1\}} (P_0(\Phi \neq 0) + P_1(\Phi \neq 1)) = 1 - \|P_0 - P_1\|_{TV} \geq 1 - \sqrt{2}d_{hel}(P_0,P_1)$.

- (Lec 19) Given sequences of tests $P_{0,n}$ versus $P_{1,n}$, we are interested in considering when the asymptotic error does not vanish, i.e. $\liminf_{n\to\infty} \inf_{\psi_n}[P_{0,n}(\psi_n \neq 0) + P_{1,n}(\psi_n \neq 1)] > 0$. This happens whenever $\liminf 1 - \sqrt{2}d_{hel}(P_{0,n},P_{1,n}) > 0$, i.e. $\limsup d_{hel}(P_{0,n},P_{1,n}) < \frac{1}{\sqrt{2}}$.

- (Lec 19) **Quadratic mean differentiability:** A family $\{P_\theta\}_{\theta\in\Theta}$ is **quadratic mean differentiable (QMD)** at $\theta \in \text{int }\Theta$ if there is a score function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$, so that

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}\dot{\ell}_\theta^T h\sqrt{p_\theta}\right)^2 d\mu = o\left(\|h\|^2\right) \qquad \text{as } h \to 0.$$

- (Lec 19) For QMD families, $P_\theta\dot{\ell}_\theta = 0$ and $P_\theta\dot{\ell}_\theta\dot{\ell}_\theta^T$ is well-defined.

- (Lec 19) Exponential families $p_\theta(x) = \exp\left[\theta^T T(x) - A(\theta)\right]$ are QMD with score $\dot{\ell}_\theta(x) = T(x) - \mathbb{E}[T(x)]$.

- (VdV Eg 7.8 p96) A location model family $\{f(x-\theta) : \theta \in \mathbb{R}\}$ is QMD if f is positive, continuously differentiable, with finite Fisher information for location $I_f = \int (f'/f)^2(x)f(x)dx$. Score function can be taken to be $-(f'/f)(x-\theta)$.

  (In particular, double exponential/Laplace location model is QMD with $\text{sign}(x-\theta)$.

- (HW9) The family $\{P_\theta\}_{\theta>0}$, where $P_\theta \sim \text{Unif}[0,\theta]$, is not QMD.

- (Lec 19, HW9) If $\{P_\theta\}$ is QMD, then $d^2_{hel}(P_{\theta+n},P_\theta) = \frac{1}{8}h^T I_\theta h + o(\|h\|^2)$, which implies that

$$\lim_{n\to\infty} d^2_{hel}(P^n_{\theta+h/\sqrt{n}},P^n_\theta) = 1 - \exp\left(-\frac{1}{8}h^T I_\theta h\right).$$

- (Lec 19, TSH Thm 12.2.3 p489, VdV Dfn 7.14) **Local asymptotic normality:** A family $\{P_\theta,n\}_{\theta\in\Theta,n\in\mathbb{N}}$ is **locally asymptotically normal (LAN)** at $\theta \in \text{int }\Theta$ with precision/information matrix $K_\theta \succeq 0$ if there exists a sequence $\Delta_n \in \mathbb{R}^d$ such that for all $h \in \mathbb{R}^d$,

$$\log\frac{dP_{\theta+h/\sqrt{n},n}}{dP_{\theta,n}} = h^T\Delta_n - \frac{1}{2}h^T K_\theta h + o_{P_{\theta,n}}(\|h\|),$$

  where $\Delta_n \xrightarrow[P_{\theta,n}]{d} \mathcal{N}(0,K_\theta)$, and $o_P(\|h\|)$ means converging in probability to 0 uniformly, if $\|h\|$ is bounded.

  - (Lec 19/20) LAN will imply contiguity (by asymptotic log normality). Le Cam's Third Lemma then implies that if $Z_n = K_\theta^{-1}\Delta_n$, then $Z_n \xrightarrow[P_{\theta+h/\sqrt{n},n}]{d} \mathcal{N}(h,K_\theta^{-1})$.
  - (Lec 19) Gaussian location family is LAN.

- (Lec 19, VdV Thm 7.2) QMD family is LAN with precision I_θ. Also, $\Delta_n = \sqrt{n} P_n \nabla \ell_\theta \xrightarrow[P_{\theta,n}]{d} \mathcal{N}(0, I_\theta)$.

From here, assume WLOG that $\theta_0 = 0$.

- (Lec 20) In an LAN family, let $Z_n = K^{-1}\Delta_n$. Then $\{Z_n\}$ is uniformly tight under $P_{h/\sqrt{n},n}$ whenever $\|h\| \leq C < \infty$.

- (Lec 20) Let $h \sim \mathcal{N}(0, \Gamma)$ with $\Gamma \succ 0$ and $Z|h \sim \mathcal{N}(Ah, \Sigma)$ with $\Sigma \succ 0$. Then

$$h|Z = z \sim \mathcal{N}\left((\Gamma^{-1} + A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}z,\ (\Gamma^{-1} + A^T\Sigma^{-1}A)^{-1}\right).$$

- (Lec 20) A function $L : \mathbb{R}^d \mapsto \mathbb{R}$ is **quasi-convex** if for all $\alpha \in \mathbb{R}$, the **α-sublevel set** $\{x : L(x) \leq \alpha\}$ is convex.

- (Lec 20) Let L be symmetric and quasi-convex. Let $A \in \mathbb{R}^{d\times k}$ and $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$\inf_{v\in\mathbb{R}^k} \mathbb{E}\left[L(AX - v)\right] = \mathbb{E}\left[L(A(X - \mu))\right] = \mathbb{E}\left[L(A\Sigma^{\frac{1}{2}}W)\right],$$

where $W \sim \mathcal{N}(0, I_k)$.

- (Lec 20, VdV Thm 8.11) **Local asymptotic minimax theorem:** Let $L : \mathbb{R}^d \mapsto \mathbb{R}$ be quasi-convex, symmetric and bounded (i.e. bowl-shaped). Let $\{P_{\theta,n}\}$ be LAN at θ_0 with precision $K_{\theta_0} \succeq 0$. Then, with $W \sim \mathcal{N}(0, I_k)$,

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{||h||\leq c,\ \theta=\theta_0+\frac{h}{\sqrt{n}}} \mathbb{E}_{P_{\theta,n}}\left[L(\sqrt{n}(\hat{\theta}_n - \theta))\right] \geq \mathbb{E}\left[L(K_{\theta_0}^{-\frac{1}{2}}W)\right].$$

(For LAN families and bowl-shaped loss, the Fisher information gives a lower bound on estimation error.)

- (VdV Lem 8.14, Thm 5.39) For most QMD families (conditions in Thm 5.39), the MLE achieves the bound in the local asymptotic minimax theorem.

13 Other Stuff

- (Lec 2) **KL divergence:** Let P and Q be distributions with densities p, q w.r.t. μ. Then we define $D_{kl}(P \parallel Q) = \int p \log\left(\frac{p}{q}\right) d\mu = \mathbb{E}_P\left[\log \frac{P}{Q}\right]$. (For discrete probability distributions, we can write it as $D_{kl}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$.)
  - $D_{kl}(P \parallel Q) \geq 0$, and $D_{kl}(P \parallel Q) = 0$ if and only if $p = q$ almost everywhere.
  - KL divergence is not symmetric.
  - If P_1 and P_2 (Q_1 and Q_2 resp.) are independent distributions with joint distribution $P(x, y) = P_1(x)P_2(y)$ (Q resp.), then $D_{kl}(P \parallel Q) = D_{kl}(P_1 \parallel Q_1) + D_{kl}(P_2 \parallel Q_2)$.
  - Pinsker's inequality: If $\lim_{n\to\infty} D_{kl}(P_n \parallel P) = 0$, then $P_n \overset{TV}{\to} P$.
  - (HW2) For an exponential family $p_\theta(x) = \exp[\langle\theta, T(x)\rangle - A(\theta)]$, $A(\theta) = \log \int \exp(\langle\theta, T(x)\rangle) d\mu(x)$. It can be computed that $D_{kl}(P_{\theta_0} \parallel P_{\theta_1}) = A(\theta_1) - A(\theta_0) - \langle\nabla A(\theta_0), \theta_1 - \theta_0\rangle$.

- (Lec 3) **Operator norm:** For $A \in \mathbb{R}^{k \times d}$, $u \in \mathbb{R}^d$, $\|A\|_{\text{op}} := \sup_{\|u\|_2 \leq 1} \|Au\|_2$.
  - The operator norm is also equal to the largest singular value of A (which are defined to be the square root of the eigenvalues of $A^T A$). When A is a real symmetric matrix, this reduces to the absolute value of the largest eigenvalue (in absolute value).
  - For any $x \in \mathbb{R}^d$, $\|Ax\|_2 \leq \|A\|_{\text{op}} \|x\|_2$.
  - $A \preceq \|A\|_{\text{op}} I$.
- (Lec 10, HW5) **Logistic regression:** $z = xy$, where $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Define $m_\theta(z) := \log\left[1 + \exp(-z^T\theta)\right] = \log\left[1 + \exp(-y\theta^T x)\right]$. This function is $\|x\|$-Lipschitz in θ.
- (TSH Lem 11.2.1 p430) **Convergence of quantiles:** Assume that F is a distribution function such that F is continuous and strictly increasing at $y = F^{-1}(1-\alpha)$.
  1. If $\{F_n\}$ is a sequence of distribution functions s.t. $F_n \Rightarrow F$, then $F_n^{-1}(1-\alpha) \to F^{-1}(1-\alpha)$.
  2. If $\{\hat{F}_n\}$ is a sequence of random distribution functions s.t. $\hat{F}_n(x) \xrightarrow{P} F(x)$ for all x which are points of continuity of F, then $\hat{F}_n^{-1}(1-\alpha) \xrightarrow{P} F^{-1}(1-\alpha)$.
- (HW2) **Property of convexity:** If function f is convex and $\nabla^2 f(\theta) \succeq \lambda I$ for all θ satisfying $\|\theta - \theta_0\| \leq c$, then
$$f(\theta) \geq f(\theta_0) + \nabla f(\theta_0)^T(\theta - \theta_0) + \frac{\lambda}{2} \min\left\{\|\theta - \theta_0\|^2, c\|\theta - \theta_0\|\right\}.$$
- (HW3) An estimator $\hat{\theta}_n$ is $\sqrt{n}$**-consistent** if $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_{P_{\theta_0}}(1)$.