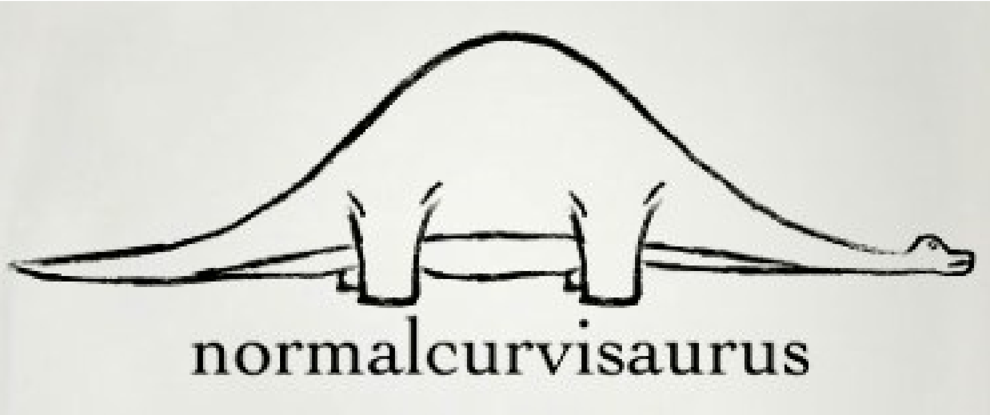
Seth Temple
sdtemple@uw.edu

Notes

STAT 580s
2020-2021

These notes are from courses taught by University of Washington Professors Alex Luedtke and Marco Carone, as well as based on van der Vaart's *Asymptotic Statistics* and Wainwright's *High-Dimensional Statistics*. My contribution is to reword and reorganize course content and to provide some supplementary material. Enjoy the cover page matter and the rest!



1 Decision Theory

We introduce four criteria to measure the performance of decision rules: Bayes risk, minimaxity, Γ-minimaxity, and admissibility.

1.1 Definitions

- An action a from an action space $\mathcal{A}$ is the realization of a probability distribution conditional on data.
- A **decision rule** $D(\cdot|X = x)$ is a probability distribution conditional on data. We commonly use deterministic decision rules, i.e. the probability distribution is degenerate.
- A loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ measures the quality of an action $a \in \mathcal{A}$ when $\theta \in \Theta$.
- Risk is a function that measures the quality of a decision rule given $\theta \in \Theta$.

$$\mathcal{R}(D, \theta) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(a, \theta) D(da|x) dP_\theta(x).$$

- A (decision) rule is **inadmissible** if there is another decision rule that has risk less than or equal to it everywhere and risk strictly less than it somewhere. A decision rule is **admissible** if it is not inadmissible.
- A rule is **minimax** if its worst case risk is the infimum of the worst case risks of various rules. That is,

$$\sup_{\theta \in \Theta} \mathcal{R}(D^\star, \theta) = \inf_{D \in \mathcal{D}} \sup_{\theta \in \Theta} \mathcal{R}(D, \theta).$$

- The Bayes risk with respect to prior Π is the expectation of a rule's risk. That is,

$$r(D, \Pi) = \int_{\Theta} \mathcal{R}(D, \theta) d\Pi(\theta).$$

- A rule is a **Bayes rule** with respect to a prior Π if it achieves the smallest Bayes risk. That is,

$$r(D^\star, \Pi) = \inf_{D \in \mathcal{D}} r(D, \Pi).$$

- A rule is unique Bayes for a prior if any other Bayes rule equals it except on null sets for all P_θ. A rule is unique minimax if any other minimax rule equals it except on null sets for all P_θ.

- An estimator $D^\star$ is Γ-minimax w.r.t. a loss function if

$$\sup_{\Pi \in \Gamma} r(D^\star, \Pi) = \inf_T \sup_{\Pi \in \Gamma} r(T, \Pi).$$

  This definition is analogous to minimax but w.r.t. Bayes risk as opposed to risk.

- The **kernel** of a posterior distribution is the function depending on the parameter, not the proportionality constant. The kernel uniquely determines the posterior distribution.

- A conjugate prior is one such that the posterior distribution is in the same family as the prior distribution.

- A prior is **least favorable** if the Bayes risk for the Bayes rule and the prior achieves the supremum over priors of Bayes risks for paired Bayes rules and priors.

- A sequence of priors is least favorable if, for all priors Π,

$$r(D_\Pi, \Pi) \leq \liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k).$$

1.2 Results

- **Finding Bayes rules by minimizing the conditional expected loss.** Suppose that $\boldsymbol{\theta} \sim \Pi$, $X|\boldsymbol{\theta} = \theta \sim P_\theta$, and the loss is nonnegative. If,

  (i) there is a rule with finite Bayes risk, and

  (ii) there exists $D_\Pi \in \mathcal{D}$ for almost all x that minimizes the conditional expected loss,

  then D_Π is a Bayes rule.

- If the loss function is convex for fixed θ, decision rules are unrestricted, the action space is convex, and there is a Bayes rule, then there is a **deterministic Bayes rule**.

- **Constant risk theorem.** If Π satisfies

$$r(D_\Pi, \Pi) = \sup_{\theta \in \Theta} \mathcal{R}(D_\Pi, \theta),$$

  then (i) D_Π is minimax, (ii) unique Bayes D_Π w.r.t. Π implies unique minimax, and (iii) Π is a least favorable prior.

- **Constant risk theorem (ii).** For a sequence of priors $\{\Pi_k\}$, if $D \in \mathcal{D}$ satisfies

$$\sup_{\theta \in \Theta} \mathcal{R}(D, \theta) = \liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k),$$

  then (i) D is minimax, and (ii) $\{\Pi_k\}$ is a least favorable prior sequence.

- **Constant Bayes risk theorem.** If $\Pi^\star \in \Gamma$ satisfies
$$r(D_{\Pi_\star}, \Pi_\star) = \sup_{\Pi \in \Gamma} r(D_{\Pi^\star}, \Pi)$$
then (i) $D_{\Pi_\star}$ is Γ-minimax, (ii) unique Bayes $D_{\Pi^\star}$ implies unique minimax, and (iii) $\Pi^\star$ is a least favorable prior.

- If a minimax estimator for a smaller model has no worse worst-case risk scenario for a large model, then the estimator is minimax for the larger model as well. (See slide 36.)

- **Some admissible estimators.** (i) Any unique Bayes rule is admissible, and (ii) any unique minimax rule is admissible. Moreover, for squared error loss, finite $r(D_\Pi, \Pi)$, and $\int P_\theta(X \in A) d\Pi(\theta) = 0$ implying $P_\theta(X \in A) = 0$ for all θ, we have that D_Π is a unique Bayes rule.

- **Stein's lemma.** Let $Y \sim N(\mu, \sigma^2)$ and let $g : \mathbb{R} \to \mathbb{R}$ be such that $\mathbb{E}[|g'(Y)|] < \infty$. Then, $\mathbb{E}[g(Y)(Y - \mu)] = \sigma^2 \mathbb{E}[g'(Y)]$. (See slides 74-76 for multivariate generalization.)

1.3 Examples

- The posterior mean is the (deterministic) Bayes rule under L_2 loss.

- The posterior median is the (deterministic) Bayes rule under L_1 loss.

- The posterior mode is the (deterministic) Bayes rule under 0-1 loss.

- The sample mean is a minimax estimator for θ in an $X \equiv (X_1, \cdots, X_n)$ iid sample of $N(\theta, \sigma^2)$ random variables.

- The sample mean is an admissible estimator for θ in the case of univariate normals with mean θ. The sample mean is inadmissible for dimension three or higher.

- The James-Stein estimator beats the sample mean when estimating a mean vector of dimension 3 or more. The positive-part James-Stein estimator beats the James-Stein estimator, so the James-Stein estimator is inadmissible too. Intuitively, the James-Stein estimator shrinks the sample mean towards zero (or some other point).
$$T^{JS}(x) = \begin{cases} (1 - \frac{(d-2)\sigma^2}{n||\bar{x}_n||^2})\bar{x}_n & \bar{x}_n \neq (0, \cdots, 0) \\ 0 & \text{otherwise} \end{cases}.$$
In some cases, we can reduce the James-Stein estimator to lower dimensions by leveraging the fact that it is a spherically symmetric estimator, where a spherically symmetric estimator is of the form $T_\tau(x) = \tau(||x||)x$. This fact solicits some geometric intuition (see slides 63-72). The shrinkage property creates bias and may be inappropriate for estimating individual means. Finally, we can motivate these estimators from an empirical Bayes perspective.

2 Large Sample Theory

These results facilitate statistical inference when samples sizes grow to infinity.

2.1 Definitions

- (convergence almost surely) $A_n \rightarrow_{a.s.} A$ if $P(\lim_{n\rightarrow\infty} ||A_n - A|| = 0) = 1$.
- (convergence in probability) $A_n \rightarrow_p A$ if, for all $\varepsilon > 0$, $P(||A_n - A|| > \varepsilon) \rightarrow 0$.
- $\mathbb{R}^d$-valued random variable A_n converges in distribution to A if, for all bounded, continuous functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

  $$\mathbb{E}[f(A_n)] \rightarrow \mathbb{E}[f(A)].$$

  This convergence is sometimes referred to as weak convergence or convergence in law.
- Uniform integrability: $\{X_n\}$ is u.i. if $\sup_n \mathbb{E}[|X_n| \cdot 1\{|X_n| \geq a\}] \rightarrow 0$. This is a condition controlling tail probabilities. (Weak convergence and u.i. imply convergence of means.)
- **Order notations.**
  - $x_n = O(r_n)$ if $\limsup |x_n/r_n| < \infty$. Equivalently, there exists $M > 0$ such that $I\{|x_n| \leq M|r_n|\} \rightarrow 1$. In layman's terms, x_n is within some multiplicative constant of r_n.
  - $x_n = o(r_n)$ if $\limsup |x_n/r_n| = 0$. Identically, for all $M > 0$, $I\{|x_n| \leq M|r_n|\} \rightarrow 1$. In layman's terms, x_n changes slower than r_n.
  - $X_n = O_P(R_n)$ if, for all $\varepsilon > 0$, there exists $M > 0$ s.t.

    $$\liminf P(||X_n|| \leq M||R_n||) > 1 - \varepsilon.$$

  - $X_n = o_P(R_n)$ if, for all $M > 0$,

    $$P(||X_n|| \leq M||R_n||) \rightarrow 1.$$

  - Stochastic and determisitic notations are equivalent when $X_n \overset{a.s.}{=} x_n$ and $R_n \overset{a.s.}{=} r_n$.
  - $X_n = o_P(1)$ if and only if $X_n \rightarrow_p 0$.
  - $X_n = O_P(1)$ is also referred to as the random sequence being uniformly tight.
  - $o_P(1)$ is related to convergence in probability whereas $O_P(1)$ is related to weak convergence.

- Read these useful properties left to right as an implication:
  (1) $X_n = o_P(R_n)$ if and only if $X_n = R_n Y_n$ for some $Y_n = o_P(1)$;
  (2) $X_n = O_P(R_n)$ if and only if $X_n = R_n Y_n$ for some $Y_n = O_P(1)$;
  (3) $o_P(1) + o_P(1) = o_P(1)$;
  (4) $o_P(1) + O_P(1) = O_P(1)$;
  (5) $O_P(1)O_P(1) = O_P(1)$;
  (6) $o_P(1)O_P(1) = o_P(1)$;
  (7) $[1 + o_P(1)]^{-1} = O_P(1)$;
  (8) $X_n = o_P(1)$ implies $X_n = O_P(1)$.

2.2 Results

- Almost sure convergence implies convergence in probability implies weak convergence.

- Let $\{A_n\}, A$, and B be defined on a common probability space. $A_n \to_{a.s.} A$ and $A_n \to_{a.s.} B$ implies $A = B$ almost surely. $A_n \to_p A$ and $A_n \to_p B$ implies $A = B$ almost surely. Similarly, if $X_n \to_d X$ and $X_n \to_d \tilde{X}$, then $X \overset{d}{=} \tilde{X}$. This juxtaposition highlights that the weak limit X is only unique up to its distribution, whereas the other convergence limits are unique.

- **Portmanteau theorem.** TFAE definitions for weak convergence $\Rightarrow$. Some of these interpretations of weak convergence are more useful than others in proving certain results. Search this list for the appropriate definition for any proof at hand.

  (i) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous f.
  (ii) $P(X_n \leq x) \to P(X \leq x)$ for all continuity points x of $P(X \leq \cdot)$.
  (iii) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, Lipschitz-continuous f.
  (iv) $\limsup \mathbb{E}[f(X_n)] \leq \mathbb{E}[f(A)]$ for every upper semicontinuous f bounded above.
  (v) $\liminf \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ for every lower semicontinuous f bounded below.
  (vi) $\limsup P(X_n \in F) \leq P(X \in F)$ for all closed sets F.
  (vii) $\liminf P(X_n \in O) \geq P(X \in O)$ for all open sets O.
  (viii) $P(X_n \in C) \to P(X \in C)$ for all continuity sets C, i.e. sets C s.t. $P(X \in \partial C) = 0$.
  (ix) $\mathbb{E}[\exp\{it^T X_n\}] \to \mathbb{E}[\exp\{it^T X\}]$ for all vectors t. (Lévy continuity)
  (x) $t^T X_n \Rightarrow t^T X$ for all vectors t. (Cramér-Wold)

- **Continuous mapping.** Let f be continuous at every point of C s.t. $P(X \in C) = 1$.

  (i) $X_n \Rightarrow X$ implies $g(X_n) \Rightarrow g(X)$;
  (ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
  (iii) $X_n \to_{a.s} X$ implies $g(X_n) \to_{a.s} g(X)$.

- **Slutsky-like lemmas.**

  (i) $X_n \Rightarrow X$ and $||X_n - Y_n|| \to_p 0$ implies $Y_n \Rightarrow X$;

  (ii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for constant c implies $(X_n, Y_n) \Rightarrow (X, c)$.

- **Slutsky's lemma.** This result provides a way to combine random vectors and random variables in the asymptote. Be careful with random vectors versus random variables in between (i) versus (ii) and (iii).

  (i) $X_n \Rightarrow X$ and $Y_n \to_p c$ for multidimensional constant c implies $X_n + Y_n \Rightarrow X + c$;

  (ii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for 1-dim constant c implies $X_n Y_n \Rightarrow cX$;

  (iii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for nonzero 1-dim constant c implies $X_n + Y_n \Rightarrow X/c$.

- **Laws of large numbers**. Let $\mathbb{E}[|X|] < \infty$. Then,

  - (weak law of large numbers) $\bar{X}_n \to_p \mathbb{E}[X]$;
  - (strong law of large numbers) $\bar{X}_n \to_{a.s} \mathbb{E}[X]$.

- **Prokhorov theorem.** This theorem relates to how $o_P(1)$ is linked with converge in probability. Here we see a relationship between $O_P(1)$ and weak convergence. The result is not quite an if and only if result. (ii) is similar to the Bolzano-Weierstrass theorem from real analysis.

  (i) $X_n \Rightarrow X$ for some X implies that $X = O_P(1)$.

  (ii) $X_n = O_P(1)$ implies that there is a subsequence $\{X_{n_i}\}$ s.t. $X_{n_i} \Rightarrow X$ for some X.

- **Central limit theorems.**

  - (vanilla univariate CLT) iid sample and finite second moment implies

    $$\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0, \sigma^2);$$

  - (vanilla multivariate CLT) iid sample and finite expected norm squared implies

    $$\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N_d(0, \sigma^2 \cdot \mathrm{Id}_d).$$

  - (Lindeberg-Feller) The setup is triangular array $\{X_{ni}\}_{i=1}^n$ with independent rows, $\mathbb{E}[X_{ni}] = \mu_{ni}$, finite $\mathrm{Var}(X_{ni}) = \sigma_{ni}^2$, $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2 > 0$, and $Y_{ni} = (X_{ni} - \mu_{ni})/\sigma_n^2$. Then, the Lindeberg condition implies $\sum_{i=1}^n Y_{ni} \Rightarrow N(0,1)$.
  - (Lindeberg) For all $\varepsilon > 0$, $\sum_{i=1}^n \mathbb{E}[Y_{ni}^2 \cdot I\{|Y_{ni}| \geq \varepsilon\}] \to 0$ as n gets large.
  - (Lyapunov) For some $\delta > 0$, $\sum_{i=1}^n \mathbb{E}[Y_{ni}^{2+\delta}] \to 0$ as n gets large. The Lyapunov condition implies the Lindeberg conditon.

- **Delta methods.**
  - (univariate) If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at ψ_0 and $r_n(\psi_n - \psi_0) \Rightarrow Z$, then
  $$r_n(f(\psi_n) - f(\psi_0)) \Rightarrow \langle Z, \nabla f(\psi_0) \rangle;$$
  - (multivariate) if $f : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable at ψ_0 and $r_n(\psi_n - \psi_0) \Rightarrow Z$, then
  $$r_n(f(\psi_n) - f(\psi_0)) \Rightarrow J_f Z,$$
  where J_f is the Jacobian with respect to function f.

2.3 Examples

- Examples
  - The vanilla univariate central limit theorem is a special case of the Lindeberg-Feller central limit theorem when we have iid samples.
  - See slides 40-44 applying Lindeberg-Feller for simple linear regression with a fixed design. This example is also discussed in Amy Willis's BIOST 533. Moreover, we can find another Lindeberg-Feller example on the BIOST 533 final exam.
  - Samples from standard multivariate normals have nice weak convergence results. See Homework 2. The crux is to decompose the random vector into polar coordinates and consider generic orthogonal transformations.
  - Estimation of relative risk using a delta method. See end of Chapter 2 slides.
- Counterexamples
  - Convergence in probability does not imply almost sure convergence. We consider a sequence of indicators that splits [0,1] into halves, then thirds, then fourths, and so on. As n gets large, this sequence of random variables converges in probability to zero. However, the indicator is triggered infinitely often, so the sequence does not converge almost surely to zero.
  - Convergence in distribution does not imply convergence in probability. We consider a sequence that converges weakly to a symmetric distribution.
  - Dependent sequences that marginally converge weakly may not jointly converge weakly. We consider sequences where the covariance between random variables alternate between -1 and 1. With independence, marginal weak convergences imply joint weak convergence.

3 M-, Z- Estimation

We introduce two paradigms for deriving estimators for a parameter θ. M for "maximum" in M-estimation involves maximizing a criterion function. Z for "zero" in Z-estimation involves finding roots of a criterion function.

3.1 Definitions

- Empirical process notation provides shorthand

$$Pf \equiv \int f(x)\, dP(x)$$

$$P_n f \equiv \frac{1}{n}\sum_{i}^{n} f(X_i)$$

- If $\phi_0 \equiv \phi(\theta_0) \in \arg\max_\phi P_0 m_\phi$, then $\phi_n \in \arg\max_\phi P_n m_\phi$ is an M-estimator. More generally, we consider M_0 and M_n that do not have to be $P_0 m_\phi$ and $P_n m_\phi$.
- If ϕ_0 is a solution to $P_0 z_\phi = \underline{0}$, then the solution ϕ_n to $P_n z_\phi = \underline{0}$ is a Z-estimator. More generally, we consider Z_0 and Z_n that do not have to be $P_0 z_\phi$ and $P_n z_\phi$.
- We call $\{m_\phi : \phi \in S \supset \text{Im}(\Phi)\}$ a P_0-GC class if $\sup_\phi |(P_n - P_0)m_\phi| = o_P(1)$
- The bracketing number measures how complex the class of functions $\mathcal{F}$ is. See slides 21-26 for an introduction with application in the Glivenko-Cantelli theorem.
- The root density $\theta \mapsto \sqrt{p_\theta}$ is **differentiable in quadratic mean** at θ if there exists a **pseudoscore** function $\dot{\ell}_\theta$ such that

$$\sup_{||h||=1} \int \left(\frac{\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta}}{\varepsilon} - \frac{h^T \dot{\ell}_\theta}{2}\sqrt{p_\theta} \right)^2 d\mu \xrightarrow{\varepsilon \to 0} 0$$

  This definition is akin to differentiability except an integral is thrown into the mix. It is exactly what we require to weaken the regularity condition of first and second order differentiability for asymptotic normality. We say a model is QMD if its root density is QMD. The pseudoscore is the score under reasonable conditions.

- The squared **Hellinger distance** is involved in QMD arguments.

$$H^2(P_\epsilon, P_0) \equiv \int \left(\sqrt{p_\epsilon} - \sqrt{p_0} \right)^2 d\mu$$

- $\mathcal{L}^2(\mu)$ contains μ-measurable functions f such that $\int f^2 d\mu$ is finite. This space is equipped with inner product and norm, so the triangle and reverse triangle inequalities hold.

3.2 Results

- Under conditions for Radon-Nikodym derivatives, the Kullback-Leibler divergence serves as an m_ϕ justifying maximum likelihood estimators as M-estimators.

- Under certain conditions, the M-estimation problem (maximizing) can be expressed as a Z-estimation problem (root-finding). On the other hand, $M_\theta(\phi) = -||Z_\theta(\phi)||$ expresses a Z-estimation problem as an M-estimation problem.

- **Uniform consistency.** Suppose that

  (i) a near-maximizer for M_n is available: ϕ_n satisfies $M_n(\phi_n) \geq \sup_\phi M_n(\phi) - o_P(1)$;

  (ii) M_n is uniformly consistent: $\sup_\phi |M_n(\phi) - M_0(\phi)| \to_p 0$; and

  (iii) ϕ_0 is well-separated: $\forall \varepsilon > 0$, $M_0(\phi_0) > \sup_{||\phi-\phi_0||>\varepsilon} M_0(\phi)$.

  Then, $\phi_n \to_p \phi_0$.

  Finding a maximum is a good way to satisfy (i). In Homework 4 Problem 1(a), we formulate a case where missing condition (iii) results in the conclusion not holding.

- **Consistency of Z-estimators in one dimension.** Let Im(Φ)$\subset \mathbb{R}$ and, for all ϕ, $Z_n(\phi) \to_p Z_0(\phi)$. One or the other or both must hold:

  (i) $\phi \mapsto Z_n(\phi)$ is continuous and has one root ϕ_n.

  (ii) $\phi \mapsto Z_n(\phi)$ is nondecreasing and there is ϕ_n such that $Z_n(\phi_n) = o_P(1)$.

  ϕ_0 such that, for all $\varepsilon > 0$, $Z_0(\phi_0 - \varepsilon) < 0 < Z_0(\phi_0 + \varepsilon)$ implies $\phi_n \to_p \phi_0$.

  This result is analogous to Homework 4 Problem 4(b).

- **Glivenko-Cantelli theorems.**

  (i) If $\mathcal{F}$ is a class of functions with finite bracketing number for all $\varepsilon > 0$, $\mathcal{F}$ is P_0-GC:

  $$||P_n - P_0||_\mathcal{F} \equiv \sup_f |(P_n - P_0)f| = o_P(1).$$

  (ii) Suppose $\mathcal{F} \equiv \{f_\phi : \phi \in K\}$ for $K \subset \mathbb{R}^d$ compact. If $\phi(x) \mapsto f_\phi(x)$ is continuous for all x and there is an envelope function F satisfying $P_0 F < \infty$ for which $\sup_\phi |f_\phi(x)| \leq F(x)$ for all x, then the bracketing number is finite for $\varepsilon > 0$.

- Under regularity conditions, we have **asymptotic normality** for Z- and M-estimators. Regularity conditions are assumptions required such that we achieve our desired result. See Chapter 3 slides 29-36 and van der Vaart Theorems 5.21 and 5.23. Below we write some of these regularity conditions for $Z-$estimation.

(i) ϕ is an open subset of $\mathbb{R}^d$;

(ii) $\mathbb{E}_0||z_{\phi_0}(X)||^2 < \infty$;

(iii) $\phi \mapsto P_{z_\phi}$ differentiable at zero ϕ_0 with nonsingular Jacobian matrix V_{ϕ_0};

(iv) There exists function G satisfying $P_0 G^2 < \infty$ so that, for all x and every ϕ and $\tilde{\phi}$ in some neighborhood of ϕ_0, $||z_\phi(x) - z_{\tilde{\phi}}(x)|| \leq ||\phi - \tilde{\phi}||G(x)$;

(v) There exists ϕ_n satisfying $P_n z_{\phi_n} = o_P(n^{-1/2})$ and $\phi_n = \phi_0 + o_P(1)$

- **Sufficient conditions for QMD.** For every θ in an open subset of $\mathbb{R}^d$, if

  (i) The root density is continuously differentiable for every x

  (ii) The information (matrix) is well-defined and continuous in θ

  then the root density is QMD and the pseudoscore is the score.

- QMD implies that the score is mean zero and that the information (matrix) exists.

- **Asymptotic normality of MLEs under QMD**. We list the regularities required.

  (i) The model is QMD at an inner point θ_0 in Θ

  (ii) There is a measurable function G with $P_0 G^2 < \infty$ such that for every θ_1 and θ_2 in a neighborhood of θ_0

  $$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \leq G(x)||\theta_1 - \theta_2||$$

  (iii) I_{θ_0} is nonsingular

  (iv) MLE is consistent

  Then,

  $$\sqrt{n}(\hat{\theta} - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0}(X_i) + o_P(1) \Rightarrow N(0, I_{\theta_0}^{-1})$$

3.3 Examples

- Method of moments estimators are Z-estimators.

- The sample median ϕ_n satisfies $(1/n)\sum_i \text{sign}(X_i - \phi) = 0$ in one dimension.

- Location-scale families are QMD under assumptions. See Homework 4 Problem 2.

- See Homework 4 Problem 3 for some general QMD models.

4 Hypothesis Testing

First, we compare the Wald, score (Rao), and likelihood ratio tests that all converge in distribution to χ^2 under regularity conditions. Second, we consider local alternatives where in sampling from the alternative we maintain desirable properties for our estimators. See Chapter 4 slides for the hypothesis testing framework.

4.1 Definitions

- The (randomized) test function $\phi_n(X)$ indicates when we reject the null.
- The **power** function $\pi_n(\theta) = \mathbb{E}_\theta[\phi_n(X)]$ measures the probability we reject the null.
- The **size** of the test is $\sup\limits_{\theta_0 \in \Theta_0} \pi_n(\theta_0)$.
- Q is absolutely continuous w.r.t. P means that $P(A) = 0$ implies $Q(A) = 0$.
- Q_n is **contiguous** w.r.t P_n means that $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$.
- A **local alternative** is some $\theta + h/\sqrt{n}$ where h describes some (small) perturbation from the null in an arbitrary direction.
- A **regular estimator** is an estimator whose sampling distribution is invariant to local perturbations of the data-generating distribution.

4.2 Results

- **Wald test.** This test rejects the null when the estimate $\hat{\psi}$ of ψ is far from zero. The test statistic $W_n \equiv n\hat{\psi}^T A_{\hat{\theta}} \hat{\psi} \Rightarrow \chi^2(m)$ under the null where m is the dimension of the space ψ lives in. This test can be easy to implement when considering many hypotheses because the we only find one MLE.
- **Likelihood ratio test.** This test rejects the null when the KL divergence is large. The test statistic $L_n \equiv 2nP_n[\ell_{\hat{\theta}} - \ell_{\hat{\theta}_0}] \Rightarrow \chi^2(m)$. This test better controls the type 1 error in small samples.
- **Score test.** This test rejects the null when the empirical mean of the score is far from zero. The test statistic $S_n \equiv Z_n(\hat{\theta}_0) I_{\hat{\theta}_0}^{-1} Z_n(\hat{\theta}_0) \Rightarrow \chi^2(m)$. This test is easy to implement because Θ_0 is often a lower-dimensional space.
- The pairwise differences between these three tests converge in probability to zero under the null hypothesis.

- **Le Cam's First Lemma**. This lemma helps us to show and characterize contiguity. TFAE:
  - (i) Q_n is contiguous w.r.t. P_n
  - (ii) $L_n \equiv \frac{dQ_n^a}{dP_n}(Z_n) \overset{P_n}{\Rightarrow} V$ along a subsequence implies that $\mathbb{E}[V] = 1$.
  - (iii) $\frac{dP_n^a}{dQ_n}(Z_n) \overset{Q_n}{\Rightarrow} U$ along a subsequence implies that $P(U > 0) = 1$.
- **Le Cam's Third Lemma.** This lemma shows us how to use contiguity when studying alternatives. Suppose that Q_n is contiguous w.r.t. P_n and $(T_n, L_n) \overset{P_n}{\Rightarrow} (T, V)$. For all measurable $A \subset \mathbb{R}^d$, let $R(A) = \mathbb{E}[I_A(T)V]$. Then R is a probability measure and $T_n \overset{Q_n}{\Rightarrow} R$.
- **Asymptotic normality of log likelihood ratio.** Suppose $\log L_n \overset{P_n}{\Rightarrow} N(\mu, \sigma^2)$. Then Q_n is contiguous w.r.t. P_n if and only if $\mu = -\sigma^2/2$. This result emphasizes that it can sometimes be useful to consider the weak limit of the log likelihood ratio.
- See van der Vaart Theorem 7.2 and Chapter 4 slides 30-40. With this theorem and Le Cam's Third Lemma, we derive results for Wald tests under local alternatives and regular estimators.
- We achieve perfect asymptotic power at fixed alternatives. That is, if we collect enough data, we will always reject a false null hypothesis.
- **Wald statistic under local alternative.** We assume the regularity conditions for the asymptotic normality of the MLE. Then, the Wald statistic W_n has a noncentral $\chi^2(m)$ weak limit where the noncentrality parameter is $h_\psi^T A_\theta h_\psi$. Because χ^2 random variables are stochastically increasing in their noncentrality parameter, the Wald test achieves non-trivial power at local alternatives.
- We skipped the section on relative efficiency. See the final Chapter 4 slides.

5 Optimality

We study when the MLE is an optimal estimator.

5.1 Results

- **Pointwise asymptotic optimality.** Suppose there is an estimator θ_n for each θ such that $\sqrt{n}(\theta_n - \theta) \overset{\theta}{\Rightarrow} Q_\theta$. For any θ there exists another estimator $\tilde{\theta}_n$ such that

$$\sqrt{n}(\tilde{\theta}_n - \theta') \overset{\theta'}{\Rightarrow} \begin{cases} Q_{\theta'} & \theta' \neq \theta \\ \delta_0 & \text{otherwise} \end{cases}$$

  That is, we can always construct a dominating estimator sequence. The above is the Hodges' estimator. See graph in Chapter 5 slides and van der Vaart Chapter 8.

- **Almost everywhere convolution.** Assume a QMD model at every θ with nonsingular information I_θ. Suppose $\sqrt{n}(\theta_n - \theta) \overset{\theta}{\Rightarrow} Q_\theta$ for every θ. Then, **for almost every** θ, there exists M_θ s.t.

$$Q_\theta \equiv Z + \epsilon$$

  where $Z \sim N(0, I_\theta^{-1})$ and $\epsilon \sim M_\theta$.

- **Convolution for regular estimators.** Let θ_n be a regular estimator sequence. Under the same conditions as the a.e. convolution theorem, **for all** θ, there exists M_θ such that

$$Q_\theta \equiv Z + \epsilon$$

  where $Z \sim N(0, I_\theta^{-1})$ and $\epsilon \sim M_\theta$.

- **Anderson's lemma.** This theorem coupled with a convolution theorem shows that the MLE is asymptotically optimal, i.e. achieves a lower bound. Let $Z \sim N(0, \Sigma)$ and $\epsilon \sim M$ be independent. If the loss L is **quasiconvex** and **centrally symmetric**, then $\mathbb{E}[L(Z)] \leq \mathbb{E}[L(Z + \epsilon)]$.

6 Minimax Lower Bounds

In nonparametric settings it is challenging to analytically derive optimal minimax risk. We look at lower and upper bounds to determine an optimal minimax rate.

6.1 Definitions

- For rule T and model $\mathcal{P}$ the **maximal risk** is $\sup_P \mathcal{R}(T, P)$. The **minimax** rule has the smallest maximal risk. Having small maximal risk means a rule will work well irrespective of the true P. It can be difficult to derive the maximal risk.
- Let $P = Q^n$ restrict us to an iid model and $\mathcal{T}_n$ denote allowable decision rules given n. A rule sequence $\{T_n\}$ is **minimax rate optimal** if

  $$\liminf_{n\to\infty} \frac{\inf_{T\in\mathcal{T}_n} \sup_Q \mathcal{R}(T, Q^n)}{\sup_Q \mathcal{R}(T_n, Q^n)} > 0$$

  Establishing rate optimality involves knowing maximal risk of T_n and minimax risk. This chapter contains methods to get lower bounds on minimax risk.
- The discrepancy $d(P_1, P_2)$ between models is $\inf_{a\in\mathcal{A}}\{L(a, P_1) + L(a, P_2)\}$.
- The testing affinity $||p_1 \wedge p_2||_1$ for models $P_1, P_2 \ll \nu$ is

  $$\int \min\left\{\frac{dP_1}{d\nu}, \frac{dP_2}{d\nu}\right\} d\nu$$

  Draw the two density curves on the same plot for further elucidation.
- The total variation distance $\text{TV}(P_1, P_2)$ is $\sup_{A\in\mathcal{A}} |P_1(A) - P_2(A)|$.
- Integrated squared error is used as the loss for regression and density problems.

  $$\text{ISE}(\hat{f}, f) = \int (\hat{f}(x) - f(x))^2 \, dx$$

- The Hölder class of functions is

  $$\mathcal{F}(\beta, L) \equiv \left\{f : |f^{(\ell)}(x_1) - f^{(\ell)}(x_2)| \leq L|x_1 - x_2|^{\beta-\ell}, \forall\, x_1, x_2\right\}$$

  where ℓ is the greatest integer less than β. So these functions are ℓ-times differentiable and satisfy some Lipschitz condition. Assuming this class is useful in smooth regression. The Lipschitz condition here, Hölder continuity, fits into a continuity heuristic.

  Continuously differentiable $\subseteq$ Lipschitz continuous $\subseteq$ Hölder continuous $\subseteq$ continuous

6.2 Results

- **Max-min inequality.** For any function $f(x, y)$ with codomain $\mathbb{R}$,

$$\inf_y \sup_x f(x, y) \geq \sup_x \inf_y f(x, y)$$

- Maximal risk is $\geq$ Bayes risk and minimax risk is $\geq$ worst-case Bayes risk.

$$\sup_P \mathcal{R}(T, P) \geq \sup_\Pi r(T, \Pi)$$

$$\inf_T \sup_P \mathcal{R}(T, P) \geq \sup_\Pi \inf_T r(T, \Pi)$$

- **Le Cam's method.** For any $P_1, P_2 \in \mathcal{P}$,

$$\inf_T \sup_P \mathcal{R}(T, P) \geq .5 \cdot d(P_1, P_2) \cdot ||p_1 \wedge p_2||_1 \geq .25 \cdot d(P_1, P_2) \cdot \exp\{-\mathrm{KL}(P_1, P_2)\}$$

  This method proposes a lower bound on minimax risk as a tradeoff between discrepancy and testing affinity or likewise discrepancy and KL divergence. Using this method yields a rate-optimal minimax lower bound for estimating a smooth density at a specific point.

- **Fano's method.** For $P_1, \cdots, P_N \in \mathcal{P}$,

$$\begin{aligned}\inf_T \sup_P \mathcal{R}(T, P) &\geq \frac{\eta}{2}\left(1 - \frac{\log 2 + N^{-1}\sum_{j=1}^N \mathrm{KL}(P_j, \bar{P})}{\log N}\right) \\ &\geq \frac{\eta}{2}\left(1 - \frac{\log 2 + \max_{j \neq k} \mathrm{KL}(P_j, P_k)}{\log N}\right)\end{aligned}$$

  where $\bar{P}$ is the uniform mixture and η is the minimum discrepancy. Using this method yields a rate-optimal minimax lower bound for smooth regression problems.

- **Eight or more lemma.** This lemma from Varshamov and Gilbert suggests that subsets of high-dimensional hypercubes exist such that the minimum Hamming distance is large. Let Ω be an m-dimensional hypercube. If $m \geq 8$, then $|\Omega| \geq 2^{m/8}$ and the minimum Hamming distance is $\geq m/8$. We take advantage of the implied inequalities in an example applying Fano's method to a smooth regression problem.

- **Pinsker's inequality.** For all distributions P_1 and P_2,

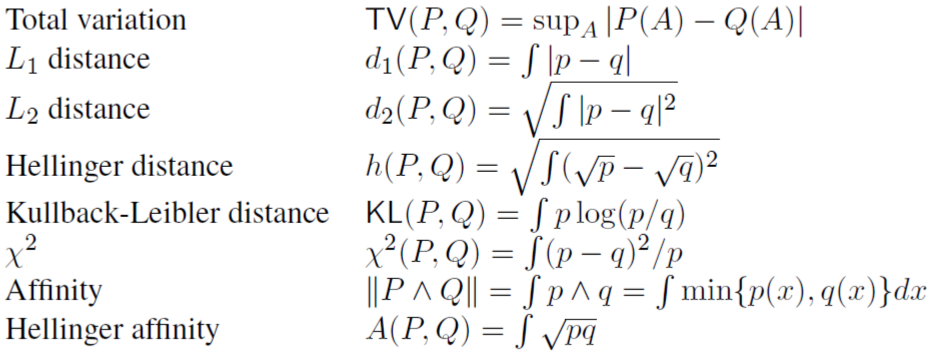
$$\mathrm{TV}(P_1, P_2) \leq \sqrt{\mathrm{KL}(P_1, P_2)/2}$$

  This belongs to a distribution distances heuristic involved in bound arguments.

$$\mathrm{TV} \leq \text{Hellinger distance} \leq \sqrt{\mathrm{KL}} \leq \sqrt{\chi^2}$$

6.3 Examples

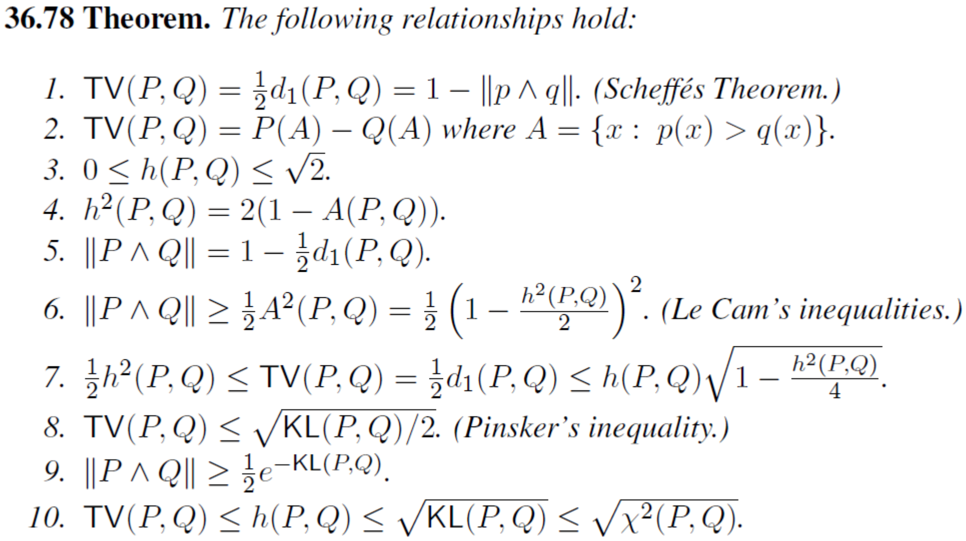
- $\text{KL}(N(\mu_1, 1), N(\mu_2, 1)) = \frac{1}{2}(\mu_1 - \mu_2)^2$
- See slides 25-36 for applying Fano's method to determine a tight lower bound for smooth regression problems. The K function on slide 27 is a bump function and $m \leq 1/h - 1$ such that bumps don't overlap and the function is infinitely differentiable (very smooth). At various points the argument is to redefine constant terms c_i and/or massage known inequalities like those from the Varshamov-Gilbert lemma.
- In this example the minimum Hamming distance appears in formulas for the discrepancy and the KL divergence. The **Hamming distance** is the number of discordant positions when comparing two byte strings.
- Applying Le Cam's two-point method:
  - For the mean θ of Rademacher random variables we derive $r_n^\star = n^{-1}$
  - Density estimation from a Lipschitz class and normal errors we derive $r_n^\star = n^{-1/3}$
- Total variation TV is difficult to work with, so we often work with KL divergence via Pinsker's inequality. When the KL divergence is large, an alternative bound provides a tighter lower bound for Le Cam's method.
- The lower bound minimax rate for estimating Hölder(β, L) density at a point is $n^{-\frac{2\beta}{2\beta+1}}$.



Total variation	$\mathsf{TV}(P, Q) = \sup_A \lvert P(A) - Q(A) \rvert$
L_1 distance	$d_1(P, Q) = \int \lvert p - q \rvert$
L_2 distance	$d_2(P, Q) = \sqrt{\int \lvert p - q \rvert^2}$
Hellinger distance	$h(P, Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
Kullback-Leibler distance	$\mathsf{KL}(P, Q) = \int p \log(p/q)$
χ^2	$\chi^2(P, Q) = \int (p - q)^2/p$
Affinity	$\lVert P \wedge Q \rVert = \int p \wedge q = \int \min\{p(x), q(x)\}dx$
Hellinger affinity	$A(P, Q) = \int \sqrt{pq}$

Figure 1: Distance metrics for probability distributions.

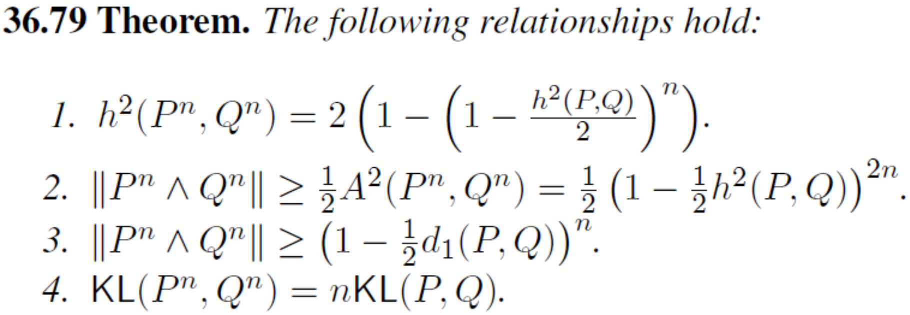
Below and on the previous page I share from Wasserman's Chapter 36 on minimax theory.



36.78 Theorem. *The following relationships hold:*

1. $\mathsf{TV}(P,Q) = \frac{1}{2} d_1(P,Q) = 1 - \|p \wedge q\|$. *(Scheffés Theorem.)*
2. $\mathsf{TV}(P,Q) = P(A) - Q(A)$ *where* $A = \{x : \ p(x) > q(x)\}$.
3. $0 \le h(P,Q) \le \sqrt{2}$.
4. $h^2(P,Q) = 2(1 - A(P,Q))$.
5. $\|P \wedge Q\| = 1 - \frac{1}{2} d_1(P,Q)$.
6. $\|P \wedge Q\| \ge \frac{1}{2} A^2(P,Q) = \frac{1}{2}\left(1 - \frac{h^2(P,Q)}{2}\right)^2$. *(Le Cam's inequalities.)*
7. $\frac{1}{2} h^2(P,Q) \le \mathsf{TV}(P,Q) = \frac{1}{2} d_1(P,Q) \le h(P,Q)\sqrt{1 - \frac{h^2(P,Q)}{4}}$.
8. $\mathsf{TV}(P,Q) \le \sqrt{\mathsf{KL}(P,Q)/2}$. *(Pinsker's inequality.)*
9. $\|P \wedge Q\| \ge \frac{1}{2} e^{-\mathsf{KL}(P,Q)}$.
10. $\mathsf{TV}(P,Q) \le h(P,Q) \le \sqrt{\mathsf{KL}(P,Q)} \le \sqrt{\chi^2(P,Q)}$.

Figure 2: Relationships between probability distances.



36.79 Theorem. *The following relationships hold:*

1. $h^2(P^n, Q^n) = 2\left(1 - \left(1 - \frac{h^2(P,Q)}{2}\right)^n\right)$.
2. $\|P^n \wedge Q^n\| \ge \frac{1}{2} A^2(P^n, Q^n) = \frac{1}{2}\left(1 - \frac{1}{2} h^2(P,Q)\right)^{2n}$.
3. $\|P^n \wedge Q^n\| \ge \left(1 - \frac{1}{2} d_1(P,Q)\right)^n$.
4. $\mathsf{KL}(P^n, Q^n) = n\mathsf{KL}(P,Q)$.

Figure 3: Relationships between probability distances.

7 Kernel Density Estimation

Kernel density estimators offer a upper bound for smooth density estimation that matches the lower bound rate, completing the argument for an optimal rate.

7.1 Definitions

- A **kernel** $K : \mathbb{R} \to \mathbb{R}$ satifies that $\int K(u)\, du = 1$. A kernel density estimator takes form

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

  where **bandwidth** h is a tuning parameter. The bandwidth controls the smoothness of the KDE, subsequently impacting the bias-variance tradeoff. Large h results in more smoothing, larger bias, and smaller variance. Small h results in less smoothing, smaller bias, and larger variance.

- Up to $s - 1$, the moments of an s^{th}-order kernel are zero and the s^{th} moment is finite.

- **Twicing** is when we combine two related KDEs to get a KDE more amenable to confidence intervals based on asymptotic normality.

$$\frac{4\hat{f}_h(x) - \hat{f}_{2h}(x)}{3}$$

7.2 Results

- The empirical cumulative distribution function is a good estimator for the CDF. However, it is a step function, so its derivative is not an ideal estimator for the density.

- Kernel density estimators for Hölder(β, L) smooth densities achieve upper bound rate

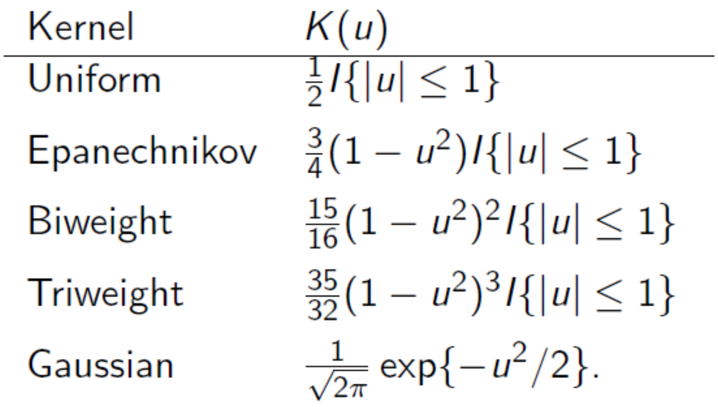
$$r_n^\star \asymp n^{-\frac{2\beta}{2\beta+1}}$$

  Thus, the rate $r_n^\star$ is an optimal rate for Hölder(β, L) density estimation. We arrive at this upper bound by setting $h_n \propto n^{-\frac{1}{2\beta+1}}$, a compromise in managing bias squared and variance. For d-dimensional densities, the rate is modified to $n^{-\frac{2\beta}{2\beta+d}}$. An example of the "curve of dimensionality".

- Kernel density derivative estimators have MSE bounded at rate $n^{-\frac{2(\beta-1)}{2\beta+1}}$

7.3 Examples

Here in two figures we see common kernels used in kernel density estimation.



Kernel	$K(u)$
Uniform	$\frac{1}{2}I\{\lvert u\rvert \leq 1\}$
Epanechnikov	$\frac{3}{4}(1-u^2)I\{\lvert u\rvert \leq 1\}$
Biweight	$\frac{15}{16}(1-u^2)^2 I\{\lvert u\rvert \leq 1\}$
Triweight	$\frac{35}{32}(1-u^2)^3 I\{\lvert u\rvert \leq 1\}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\{-u^2/2\}.$

Figure 4: Functional forms for common kernels.

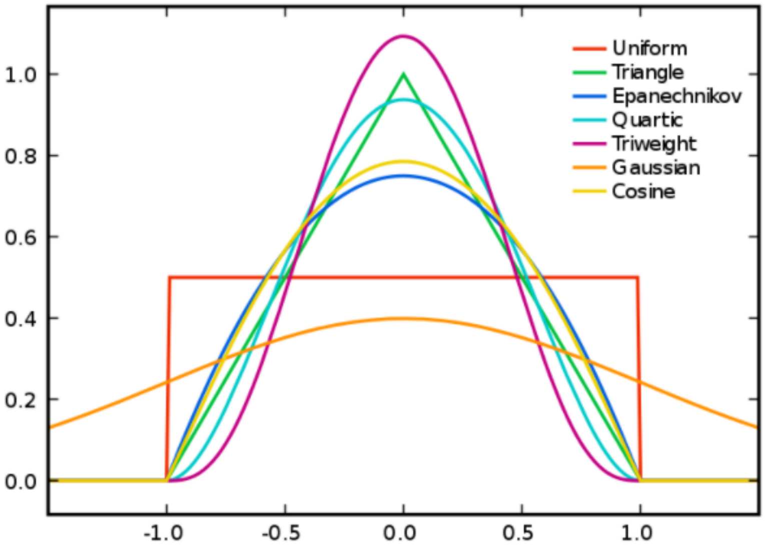


Figure 5: Graphs for common kernels.

8 Concentration Inequalities

Here we study assurances that finite samples are concentrated about their mean.

8.1 Definitions

- A random variable X is **sub-Gaussian** if its tails are lighter than Gaussian tails. Parameterized with σ^2, for all $\lambda \in \mathbb{R}$ it holds that
$$\log M_{X-\mathbb{E}[X]}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$
- A nonnegative random variable X is **subexponential** if its tails are lighter than exponential tails. Parameterized with σ^2 and b, for all $|\lambda| < 1/b$ it holds that
$$\log M_{X-\mathbb{E}[X]}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$
- A **Rademacher** random variable puts 1/2 weight on 1 and 1/2 weight on -1. It is like a Bernoulli random variable but with support $\{-1, 1\}$.
- A function f satisfies the **bounded differences property** (BDP) if for all i there is some finite c_i such that
$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_{i'}, x_{i+1}, \ldots, x_n)| \leq c_i$$
This property says that f cannot rely too much on any single input.

8.2 Results

- **Markov's inequality.** Let X be a nonnegative random variable with finite expected value. For $t > 0$,
$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$
- **Markov's corollary.** Let f be a nondecreasing function on nonnegative values and $\mathbb{E}[f(|X - \mathbb{E}[X]|)]$ finite. Extending Markov's inequality, for $t > 0$,
$$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[f(|X - \mathbb{E}[X]|)]}{f(t)}$$
This corollary shows us how to build on Markov's inequality to generate bounds on higher moments, e.g. Chebyshev's inequality and the Chernoff bound.

- **Chernoff bound.** Suppose X has moment generating function in a neighborhood of zero. For $t > 0$ and $\lambda \in (0, b]$,

$$P(X - \mathbb{E}[X] \geq t) \leq \frac{\mathbb{E}[\exp\{\lambda(X - \mathbb{E}[X])\}]}{e^{\lambda t}} = \frac{M_{X-\mathbb{E}[X]}(\lambda)}{e^{\lambda t}}$$

$$P(X - \mathbb{E}[X] \geq t) \leq \inf_{\lambda > 0} \frac{M_{X-\mathbb{E}[X]}(\lambda)}{e^{\lambda t}}$$

- **Additivity of subexponentials and sub-Gaussians.** This property is inherited as a property of moment-generating functions when we have independent random variables. See discussion on page 29 of Wainwright (2019). For independent subexponential $X_1, \ldots, X_n$ with parameters (σ_i^2, b_i), the sum of these is subexponential with parameters $(\sum_{i=1}^n \sigma_i^2, \max_i b_i)$.

- **Hoeffding's inequality.** This inequality applies to random variables with bounded support $[a, b]$. These random variables are sub-Gaussian with $\sigma^2 = (b - a)^2/4$. The second statement uses independence to extend the result to a sample mean.

$$\log P(X - \mathbb{E}[X] \geq t) \leq -\frac{2t^2}{(b-a)^2}$$

$$\log P(\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq t) \leq -\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}$$

- **Tail bound for subexponentials.** If X is subexponential with parameters (σ^2, b),

$$\log P(X - \mu \geq t) \leq \begin{cases} -\frac{t^2}{2\sigma^2}, & 0 \leq t \leq \sigma^2/b \\ -\frac{t}{2b}, & t > \sigma^2/b \end{cases}$$

  This result suggests that the (right) tail probability in the exponent is quadratic in t for small t and linear in t for large t. See an alternative presentation with proof as Proposition 2.9 from Wainwright (2019).

- **Bernstein's inequality.** Let X have finite variance σ^2 and be bounded s.t. $|X - \mu| \leq b$. For all $t > 0$ we have

$$P(X \geq \mu + t) \leq \exp\left(- \frac{t^2}{2(\sigma^2 + bt)} \right)$$

  If we have an independent collection of random variables so bounded and with possibly unique mean and finite variances, then for all $t > 0$

$$P(\bar{X}_n \geq \mathbb{E}[\bar{X}_n] + t) \leq \exp\left(- \frac{t^2}{2(bt + \sum_{i=1}^n \sigma_i^2/n)} \right)$$

  Bernstein's may be tighter than Hoeffding's when variances are small.

- Bennet's inequality. Given independent $X_1, \dots, X_n$ with zero mean, $X_i \in [-b, b]$, and variances σ_i^2,

$$P(\bar{X} \geq t) \leq \exp\Bigg(- \frac{n\bar{\sigma}^2}{b^2} h\bigg(\frac{bt}{\bar{\sigma}^2}\bigg)\Bigg)$$

  where $\bar{\sigma}^2 = 1/n \sum_{i=1}^n \sigma_i^2$ and $h(y) = (1+y)\log(1+y) - y$ for $y \geq 0$. This inequality is at least as tight as Bernstein's. See Exercise 2.7 in Wainwright for a walkthrough.

- **Efron-Stein inequality.** Let $f : \mathcal{X} \to \mathbb{R}$, and assume independent random variables.

$$\begin{aligned} \mathrm{Var}(f(\mathbf{X})) &\leq \frac{1}{2} \sum_{i=1}^n \mathbb{E}[(f(\mathbf{X}) - f(\mathbf{X}^{(i)}))^2] \\ &= \sum_{i=1}^n \mathbb{E}[(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})|X^{(i)}])^2] \end{aligned}$$

  Further assumptions on f enable nicer forms on the bound. See homework 2. Proving this involves a novel strategy.

- **Bounded differences inequality.** Let $X = (X_1, \dots, X_n)$ be an independent collection of random variables and f satisfy BDP with bounds $c_1, \dots, c_n$. For $t > 0$ and $\mathbb{E}[|f(X)|]$ finite,

$$P(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2\exp\Bigg\{ - \frac{2t^2}{\sum_{i=1}^n c_i^2} \Bigg\}$$

  This inequality is referred to as McDiarmid's inequality. Proving it involves a novel telescoping argument and the Azuma-Hoeffding inequality. See slides 34-39.

- **Martingale differences inequality**. Consider the martingale difference sequence:

$$\begin{aligned} D_1 &= \mathbb{E}[f(X)|X_1] - \mathbb{E}[f(X)] \\ D_2 &= \mathbb{E}[f(X)|X_1, X_2] - \mathbb{E}[f(X)|X_1] \\ &\vdots \\ D_n &= \mathbb{E}[f(X)|X_1, \dots, X_n] - \mathbb{E}[f(X)|X_1, \dots, X_{n-1}] \end{aligned}$$

  Further assume the conditions of the bounded differences inequality. Then,

$$P\Bigg(\bigg| \sum_{i=1}^n D_i \bigg| \geq t \Bigg) \leq 2\exp\Bigg\{ - \frac{2t^2}{\sum_{i=1}^n c_i^2} \Bigg\}$$

  This inequality is referred to as the Azuma-Hoeffding inequality.

8.3 Examples

- Sub-Gaussian random variables.
  - Gaussian
  - Beta
  - Uniform
  - Any bounded random variable
- **Not** sub-Gaussian random variables.
  - Double-exponential
  - Student t
  - Cauchy
- Subexponential random variables.
  - Exponential
  - Gaussian
  - χ^2
- **Not** subexponential random variables.
  - Log-Normal
  - Weibull
  - Pareto
  - **Not** sub-Gaussian random variables
- See Examples 2.11 and 2.12 in Wainwright (2019). Together these examples show how to reduce dimension based on a random projection while maintaining certain guarantees about the dimension reduction. This is done by leveraging properties of subexponential χ^2 random variables.
- See Slides 14-16 from Chapter 3 for a discussion of the Chernoff bound for normal random variables being in some sense aymptotically tight. Additionally, this example shows how to derive MGFs using a change of variables.

9 Empirical Risk Minimization (ERM)

We establish tools to bound the regret in ERM by bounding empirical process terms, e.g. bounding the Rademacher complexity.

9.1 Definitions

- For a set Θ, an empirical risk minimizer $\hat{\theta}$ is such that $P\ell(\hat{\theta})$ is close to $\inf_{\theta\in\Theta} P\ell(\theta)$, denoting θ_0 here as the arg min. In decision theory, we talk about θ as a parameter that characterizes a parametric family. As a result, we look explicitly at the discrepancy between $\hat{\theta}$ and θ. Here we instead look at differences in expectation, introducing the term **regret**:
$$\text{Reg}(\hat{\theta}) = P(\ell(\hat{\theta})) - P(\ell(\theta_0))$$

- A **ghost sample** is an independent (copy) sample $X_1', \ldots, X_n'$ used in Rademacher symmetrization arguments. See Chapter 4 Part 1 Slides 13-18 for this technique.

- Let $\varepsilon_1, \ldots, \varepsilon_n$ be Rademacher random variables, and define the **Rademacher process**
$$R_n := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)$$
We denote $||R_n||_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |R_n(f)|$ and call $\mathbb{E}\,||R_n||_{\mathcal{F}}$ the **Rademacher complexity**.

- The growth function or **shattering** number measures the richness of the class $\mathcal{F}$
$$\Pi_{\mathcal{F}}(n) := \sup_{x_1,\ldots,x_n} |\mathcal{F}_{x_1,\ldots,x_n}|$$
We discuss shattering in the context of function classes or sets. We say n points are shattered by $\mathcal{F}$ if $\Pi_{\mathcal{F}}(n) = 2^n$. Below are some properties of growth functions.
  - $\Pi_{\mathcal{A}}(n+m) \leq \Pi_{\mathcal{A}}(n)\,\Pi_{\mathcal{A}}(m)$
  - $\Pi_{\mathcal{A}\cup\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n) + \Pi_{\mathcal{B}}(n)$
  - $\Pi_{A\cup B: A\in\mathcal{A}, B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\,\Pi_{\mathcal{B}}(n)$
  - $\Pi_{A\cap B: A\in\mathcal{A}, B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\,\Pi_{\mathcal{B}}(n)$

- The **Vapnik-Chervonenkis (VC) dimension** is the largest n such that a function class still shatters n points. The **VC index** is one plus the VC dimension, i.e. the first n such that a function class can't shatter n points. Write VC($\mathcal{F}$) as VC dimension of $\mathcal{F}$. We say a function class $\mathcal{F}$ (or set $\mathcal{A}$) is VC if it has finite VC dimension. VC theory is for classification problems.

- For $r \geq 1$, $L^r(P)$ is the space of functions f s.t. $||f||_{L^r(P)} := (\int |f(x)|^r \, dP(x))^{1/r} < \infty$. The sup norm or uniform norm is $\sup_{f \in \mathcal{F}} |f|$.

- A Glivenko-Cantelli (GC) theorem is one where the implication is $||P_n - P||_{\mathcal{F}} = o_P(1)$.

- A bracket $[\ell, u]$ is the set of $f \in \mathcal{F}$ s.t. $\ell \leq f \leq u$ pointwise. An ϵ-bracket is a bracket $[\ell, u]$ satisfying $||u - \ell||_{L^r(P)} \leq \epsilon$. The **bracketing number** $N_{[]}(\epsilon, \mathcal{F}, L^r(P))$ is the minimum cardinality of ϵ-brackets required to cover function class $\mathcal{F}$.

- A **pseudometric** d, related to a metric, has ii) symmetry and iii) triangle inequality, but does not have i) the identity of indiscernibles. $d(x, y) = 0$ does not imply $x = y$. This is a more flexible quantity for discussing functions in $L^r(P)$ space.

- A subset T_1 of a pseudometric space T is an **ϵ-cover** if for each $\theta \in T$ there is $\theta_1 \in T_1$ s.t. $d(\theta_1, \theta) \leq \epsilon$. Visually, T is a subset of a union of ϵ-balls centered at point in T_1. The **ϵ-covering number** $N(\epsilon, T, d)$ is the minimum cardinality among possible ϵ-covers. The logarithm of the ϵ-covering number is refered to as the **metric entropy**. (We say T is totally bounded if for all positive ϵ there exists a finite ϵ-cover.)

- A subset T_1 of a pseudometric space T is an **ϵ-packing** if $d(\theta, \theta') > \epsilon$ for each pair $\theta, \theta' \in T_1$. Visually, ϵ-balls centered at points in T_1 do not overlap. The **ϵ-packing number** $M(\epsilon, T, d)$ is the maximum cardinality among possible ϵ-packings.

- A stochastic process $\{X_\theta : \theta \in T\}$ is a collection of (indexed) random variables. It is zero-mean if $\mathbb{E}[X_\theta] = 0$ for all indices θ. It is **sub-Gaussian** w.r.t. d if, for all pairs $\theta, \theta' \in T$ and for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp\{\lambda(X_\theta - X_{\theta'})\}] \leq \exp\left(\frac{\lambda^2 d(\theta, \theta')}{2}\right)$$

  This generalizes sub-Gaussianity from Section 8 to stochastic processes.

- The **canonical Rademacher process** is a zero-mean, sub-Gaussian random process:

$$X_\theta = \sum_{i=1}^{n} \theta_i \varepsilon_i = \langle \theta, \varepsilon \rangle$$

  The canonical Gaussian process is a defined in the same fashion with $\varepsilon_1, \ldots, \varepsilon_n$ being standard normals instead of Rademachers.

- The diameter D of a (pseudometric) space T is $\sup_{\theta, \theta'} d(\theta, \theta')$.

- An envelope function F for $f \in \mathcal{F}$ is s.t. $|f| \leq F$ pointwise.

9.2 Results

- **Regret bound.** See slides 9 and 10 for explanation of notation.

$$\begin{aligned} 0 &\le \mathrm{Reg}(\hat{\theta}) \\ &\le (P_n - P)(\ell(\theta_0) - \ell(\hat{\theta})) \\ &\le 2 \sup_{f \in \mathcal{F}} |(P - n - P)f| \\ &= 2\,||P_n - P||_{\mathcal{F}} \end{aligned}$$

  Often we bound $||P_n - P||_{\mathcal{F}}$ or $||R_n||_{\mathcal{F}}$, but this loosening may be suboptimal. In Homework 5(c) we attained a tighter bounded working with $(P_n - P)(\ell(\theta_0) - \ell(\hat{\theta}))$.

- **Symmetrization bound.** We relate $||P_n - P||_{\mathcal{F}}$ to the Rademacher process via a symmetrization argument.

$$\mathbb{E}\,||P_n - P||_{\mathcal{F}} \le 2\,\mathbb{E}\,||R_n||_{\mathcal{F}}$$

  Here the argument precedes by introducting the ghost sample, recognizing that $|\varepsilon_i| = 1$, and applying the triangle inequality. A related lower bound is available by desymmetrization (similar strategy).

- Tight control using Rademacher complexity. If $\mathcal{F}$ is a class of [0,1]-valued functions,

$$\frac{1}{2}\mathbb{E}\,||R_n||_{\mathcal{F}} - \sqrt{\frac{\log 2}{2n}} \le \mathbb{E}\,||P_n - P||_{\mathcal{F}} \le 2\,\mathbb{E}\,||R_n||_{\mathcal{F}}$$

  With probability $1 - 2\exp(-2nt^2)$, for all $t > 0$

$$\mathbb{E}\,||P_n - P||_{\mathcal{F}} - t \le ||P_n - P||_{\mathcal{F}} \le \mathbb{E}\,||P_n - P||_{\mathcal{F}} + t$$

  The second result uses the first result and applies McDiarmids' BDP inequality.

- Counting operations. For a family of boolean-valued functions

$$\mathcal{F} = \{x \mapsto f(x, \theta : \theta \in \mathbb{R}^p)\},$$

  where each $f : \mathbb{R}^m \times \mathbb{R}^p \to \{0, 1\}$, suppose f is computed in no more than t arithmetic or comparison operations. Then, $\mathrm{VC}(\mathcal{F}) \le 4p(t + 2)$.

- **Finite class lemma.** If $\mathcal{F}$ is a function class s.t. $|f(x)| \le 1$,

$$\mathbb{E}\,||R_n||_{\mathcal{F}} \le \sqrt{\frac{2\log(2\,\mathbb{E}\,|\mathcal{F}_{X_1^n}|)}{n}}$$

  This pivotal result provides a bound for the Rademacher complexity when $\mathcal{F}$ is a bounded function class. The proof strategy is to condition on X_1^n, use Jensen's and sub-Gaussianity to loosen the bound, and then integrate over X_1^n. When $\mathcal{F}$ is VC,

$$\mathbb{E}\,||P_n - P||_{\mathcal{F}} \le 2\,\mathbb{E}\,||R_n||_{\mathcal{F}} \le 2\sqrt{2\log(2\Pi_{\mathcal{F}}(n))/n}$$

- **Generalized finite class lemma.** If $\{X_\theta : \theta \in T\}$ is sub-Gaussian w.r.t. d and $A \subseteq T \times T$,
$$\mathbb{E} \max_{(\theta,\theta')\in A} (X_\theta - X_{\theta'}) \leq \sqrt{2 \log |A|} \cdot \max_{(\theta,\theta')\in A} d(\theta,\theta')$$

- **Sauer's lemma.** If $\text{VC}(\mathcal{F}) \leq d$, then we have two bounds on the growth function.
$$\Pi_{\mathcal{F}}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$$
$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n, & n \leq d \\ (e^1 n/d)^d, & n > d \end{cases}$$
Past the VC dimension, we're only polynomial in n. When $\text{VC}(\mathcal{F}) \leq d < n$,
$$\mathbb{E}\, ||P_n - P||_{\mathcal{F}} \leq 2\, \mathbb{E}\, ||R_n||_{\mathcal{F}} \leq 2\sqrt{(2\log(2) + 2d\log(e^1 n/d))/n}$$

- **Standard Glivenko-Cantelli.** If $N_{[]}(\epsilon, \mathcal{F}, L^1(P))$ finite for every positive ϵ, then
$$||P_n - P||_{\mathcal{F}} = o_P(1)$$

- More general GC. If the sup metric entropy over all discrete distributions Q is finite for all positive ϵ and $f \in \mathcal{F}$ have range in $[-M, M]$,
$$||P_n - P||_{\mathcal{F}} = o_P(1)$$

- **Covering versus packing.** For all $\epsilon > 0$,
$$M(2\epsilon, T, d) \leq N(\epsilon, T, d) \leq M(\epsilon, T, d)$$

- Bracketing versus covering. Let $\mathcal{F}$ be a subset of some $L^r(P)$ space. For $\epsilon > 0$,
$$N_{[]}(2\epsilon, \mathcal{F}, L^r(P)) \leq N(\epsilon, \mathcal{F}, ||\cdot||_\infty)$$

- **Discretization bound.** Let $\{X_\theta : \theta \in T\}$ be a zero-mean, sub-Gaussian process, and let D be the diameter of T. For all positive ϵ,
$$\mathbb{E} \sup_{\theta\in T} X_\theta \leq 2\, \mathbb{E} \sup_{d(\theta,\theta')\leq\epsilon} (X_\theta - X_{\theta'}) + 2D\sqrt{\log N(\epsilon, T, d)}$$
We achieve this by a ± trick, loosening with sup, and applying a finite class lemma.

- **Chaining bound.** Same setup as for the discretization bound. For all positive ϵ,

$$\mathbb{E}\sup_{\theta\in T} X_\theta \leq \mathbb{E}\sup_{d(\theta,\theta')\leq\epsilon}(X_\theta - X_{\theta'}) + 8\int_{\epsilon/2}^{D}\sqrt{\log(N(\delta,T,d))}\,d\delta$$

For separable stochastic processes,

$$\mathbb{E}\sup_{\theta\in T} X_\theta \leq 8\int_0^D \sqrt{\log N(\delta,T,d)}\,d\delta$$

Alex offered an extra lecture on proving the separable case. We achieve this by iteratively applying a finite class lemma in a clever way. This bound is generally tighter than the discretization bound.

- **Chaining to control Rademacher complexity.** If $\mathcal{F}$ is closed under negations,

$$\mathbb{E}\,||R_n||_{\mathcal{F}} \leq 8n^{-1/2}\,\mathbb{E}\int_0^\infty \sqrt{\log N(\delta,\mathcal{F},L^2(P_n))}\,d\delta \leq 8n^{-1/2}\sup_Q\int_0^\infty \sqrt{\log N(\delta,\mathcal{F},L^2(Q))}\,d\delta$$

- Bracketing integral bound. There is a universal positive constant C s.t for any class $\mathcal{F}$ with envelope function F,

$$\mathbb{E}\,||P_n - P||_{\mathcal{F}} \leq Cn^{-1/2}\,||F||_{L^2(P)}\int_0^1 \sqrt{\log N_{[]}(\delta||F||_{L^2(P)},\mathcal{F},L^2(P))}\,d\delta$$

9.3 Examples

- Vapnik-Chervonenkis
  - Dimension/index (Homework 3 Problem 4)
    - $\{(-\infty,b] : b\in\mathbb{R}\}$ has VC index 2.
    - $\{(-\infty,b_1]\times(-\infty,b_d] : b_1,\ldots,b_d\in\mathbb{R}\}$ has VC index $d+1$
    - $\{(a_1,b_1]\times(a_1,,b_d] : a_1,\ldots,a_d,b_1,\ldots,b_d\in\mathbb{R}\}$ has VC index $2d+1$
    - $\{(a,b] : a,b\in\mathbb{R}\}$ has VC index 3.
    - $\{(a_1,b_1]\times(a_2,b_2] : (a_1,a_2),(b_1,b_2)\in\mathbb{R}^2\}$ has VC index 5.
    - $\{x\mapsto g(x-\theta) : \theta\in\mathbb{R}, g:\mathbb{R}\to\mathbb{R}\text{ monotone}\}$ has VC index 2.
    - Collection of convex sets in $\mathbb{R}^2$ has VC index ∞.
    - $\{x\in\mathbb{R}^2 : ||x-a||_2\leq b\}$ has VC index 4.
  - Applying the "counting operations" lemma
    - Linear threshold class is VC
    - Neural network classifiers are VC and we can attain probabilistic guarantees for regret (Homework 4 Problem 2)

- Permanence properties of VC classes (Homework 4 Problem 1)
    - $\{\{x : f(x) > 0\} : f \in \mathcal{F}\}$
    - $\{\{x \mapsto f(x) + g(x)\} : f \in \mathcal{F}, g \text{ fixed}\}$
    - $\{\{x \mapsto f(x)g(x)\} : f \in \mathcal{F}, g \text{ fixed}\}$

- Bracketing
    - Proving the standard GC theorem (Chapter 4 Part 2 Slides 5-8)
    - Homework 4 Problem 3 (very hard, see solutions)
    - Sobolev classes (not covered, Chapter 4 Part 2 Slides 58)
    - Various examples in vdV textbook (Chapter 19)

- ϵ-covering
    - For a bounded r-dimensional set $N(\epsilon) \asymp \epsilon^{-r}$ for any ℓ_p metric
    - $(r/\epsilon)^d \leq N(\epsilon, B^d(0, r), \ell_p) \leq (2r/\epsilon + 1)^d$
    - $N(\epsilon, \mathcal{F}, || \cdot ||_{\mathcal{F}}) \leq N(\epsilon/L, B, || \cdot ||_b)$ for $\mathcal{F} := \{x \mapsto f(x, \beta : \beta \in B)\}$ with an L-Lipschitz condition (Chapter 4 Part 2 Slides 17-18)
    - $\log N(\epsilon, \mathcal{F}, || \cdot ||_\infty) = \Theta(L/\epsilon)$ for L-Lipschitz $[0, 1] \to [0, 1]$ functions
    - $\log N(\epsilon, \mathcal{F}, || \cdot ||_\infty) = \Theta((L/\epsilon)^d)$ for L-Lipschitz $[0, 1]^d \to [0, 1]$ functions
    - $\log N(\epsilon, T, \ell_2) = \log N(\epsilon^{-1/2}, \mathcal{F} \cup -\mathcal{F}, L^2(P_n))$
    - Homework 4 Problem 4 (very hard, see solutions)
    - Various examples in Wainwright textbook (Chapters 4 and 5)

- ϵ-packing
    - $M(\epsilon, B^d(0, r), \ell_p) \leq (3r/\epsilon)^d$
    - Any ϵ-covering example can be translated to be an ϵ-packing example

- The one-step discretization bound for the canonical Rademacher process provides

$$\mathbb{E} \sup_{\theta \in T} X_\theta \leq 2n^{1/2}\delta + 2D\sqrt{\log N(\delta, T, \ell_2)}$$

- The one-step discretization bound for Rademacher complexity provides

$$\mathbb{E}\, ||R_n||_{\mathcal{F}} \leq 2\delta + 2\,\mathbb{E}[D_{Z_1^n}]n^{-1} \sup_Q \sqrt{\log 2N(\delta, \mathcal{F}, L^2(Q))}$$

Miscellaneous

- **Differentiability in higher dimensions.** Below are two representations of differentiability at ψ_0. The first representation is especially useful for proving delta methods. For all $\varepsilon > 0$,
$$\lim_{\varepsilon \to 0} \sup_{||h||=1} \frac{|f(\psi_0 + \varepsilon h) - f(\psi) - \varepsilon\langle h, \nabla f(\psi_0)\rangle|}{\varepsilon} \to 0;$$
$$\lim_{h \to 0} \frac{||f(\psi_0 + h) - f(\psi) - J_f(h)||}{||h||} \to 0.$$
Read more about the second representation at the Wikipedia article for differentiable functions in higher dimensions. A sufficient condition for differentiability is that partial derivatives exist and the linear map J_f is the Jacobian matrix. In class, we claim this sufficient condition as that f is partially differentiable in a neighborhood around ψ_0 and the partial derivatives are continuous at ψ_0. Lastly, note that h in the two representations are different!

- Find a Lipschitz constant by finding the maximum of the norm of the gradient.

- **Mean value theorem.** For a function continuous on $[a, b]$ and differentiable on (a, b) there is some c in the open interval such that the tangent line $f'(c)$ equals the secant line connecting a and b.

- **Extreme value theorem.** A continuous function on a closed interval attains its maximum and minimum values. Intermittently we assert that a extremum is at least attained based on this result.

- **Fundamental theorems of calculus.** These results links derivatives with integrals.
$$f(b) - f(a) = \int_a^b f'(x)\, dx$$
$$f(x) = \int_a^x f(t)\, dt \implies f'(x) = f(x)$$

- **Taylor series.** We approximate differentiable functions by the following and the remainder vanishes in the asymptote.
$$f(x) = \frac{f^{(0)}(a)}{0!}(x-a)^0 + \frac{f^{(1)}(a)}{1!}(x-a)^1 + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \text{remainder}$$

- **Lagrange remainder.** Combining Taylor series with the mean value theorem we arrive at an alternative representation for ℓ-differentiable functions.
$$f(x) = \frac{f^{(0)}(a)}{0!}(x-a)^0 + \frac{f^{(1)}(a)}{1!}(x-a)^1 + \cdots + \frac{f^{(\ell-1)}(a)}{(\ell-1)!}(x-a)^{\ell-1} + \frac{f^{(\ell)}(\tilde{x}_a)}{\ell!}(\tilde{x}_a)^\ell$$
Here $\tilde{x}_a$ is some point in the interval between x and a. Final term is Lagrange remainder.

- **Properties of convex functions.**
  - $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$ for $t \in [0,1]$.
  - With differentiability added, the function is continuously differentiable.
  - Differentiable on an $[a,b]$ implies the derivative is nondecreasing on $[a,b]$.
  - Differentiable on $[a,b]$ implies

    $$f(x) \geq f(y) + f'(y)(x-y) \qquad x,y \in [a,b]$$

- Use this equality $\inf(-x_n) = -\sup x_n$ to redefine the $X_n = O_P(R_n)$ notation.

- $1+x \leq e^x$ is a useful upper bound inequality to convert to an exponential argument

- $\log(1+x) \leq x$ for $x \geq 0$ is a useful upper bound inequality to convert to

- $(e^x - x - 1)/x^2$ is a nondecreasing function

- Some useful Taylor series
  - $e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$
  - $\log(1+x) = 0 + \frac{x^1}{1!} - \frac{x^2}{2!} + \frac{x^3}{3!} - \frac{x^4}{4!} \pm \dots$
  - $\frac{1}{1-x} = 1 + x^1 + x^2 + x^3 + \dots$

- **Reverse triangle inequality.**

  $$\Big| ||x||_2 - ||y||_2 \Big| \leq ||x-y||_2$$

- **Bolzano-Weierstrass.** Every bounded sequence has a convergent subsequence.

- **Semi-continuity.** See graphs under Examples.
  - A function f is lower-semicontinuous at x_0 if for all $\varepsilon > 0$ there is a neighborhood around x_0 such that $f(x) \geq f(x_0) - \varepsilon$
  - A function f is upper-semicontinuous at x_0 if for all $\varepsilon > 0$ there is a neighborhood around x_0 such that $f(x) \leq f(x_0) + \varepsilon$.

- **Fisher-Cramér.** This result from MDP's STAT 513 says that the MLE is asymptotically consistant and normal if we have a regularity condition the depends of first and second derivatives of the log likelihood. Important estimators like medians may not fit into this framework. We weaken this assumption using QMD.

- **Binomial theorem.**

  $$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Notes

- We say that X is stochastically larger than Y if $P(X \leq x) \leq P(Y \leq x)$.
- See this for functions of bounded variation and how to (Jordan) decompose such.
- For concave functions, the inequality sign in Jensen's switches.
- $(y-a)^2 - (y-b)^2 = (a-b)(a+b-2y)$ we use studying $\text{Reg}(\hat{\theta})$ for squared loss.
- Symbol $\asymp$ refers to order of magnitude.
- Symbol $\lesssim$ refers to $\leq$ up to constants.
- Tight bound $\Theta(\cdot)$ is explained here.
- $\text{vol}(B^d(0,\epsilon)) \leq (\epsilon/r)^d \text{vol}(B^d(0,r))$