Infotheory for Statistics and Learning

Lecture 13

- Method of types in action[1]
  - Recap
  - Conditional limit theorem
  - Hypothesis testing
  - Neyman Pearson's Lemma
  - Stein's Lemma
  - Chernoff information

[1] based on material in [CS] and [CT].

Recap: Sanov's Theorem

- Let $x^n = (x_1, x_2 \ldots, x_n) \in \mathcal{A}^n$ denote a sequence of length n defined on finite set $\mathcal{A} = \{a_1, a_2, \ldots a_M\}$ with *empirical distribution* $\hat{P}_{x^n}$, which is also the *type* of the sequence.
- The probability that a sequence drawn iid $\sim Q$ will have type $\hat{P}_{X^n}$ depends exponentially on the distance $D(\hat{P}_{X^n}||Q) \cdot n$.
- Sanov's Theorem asks for the probability that the type $\hat{P}_{X^n}$ will be in a set $\mathcal{E} \subset \mathcal{P}$. We again observe an exponential decay rate, but the decay depends on the smallest distance between Q and distributions $P \in \mathcal{E}$, i.e. $D(\mathcal{E}||Q) = \inf_{P \in \mathcal{E}} D(P||Q)$.

Sanov's Theorem: Let $\mathcal{E}$ be a set of distribution whose closure is equal to it closure of its interior. Then for the empirical distribution $\hat{P}_{x^n}$ of a sample sequence iid of strictly positive distribution Q on $\mathcal{A}$ we have

$$-\frac{1}{n} \log \mathrm{Prob}\{\hat{P}_{X^n} \in \mathcal{E}\} \overset{n\to\infty}{\longrightarrow} D(\mathcal{E}||Q).$$

Recap: Proof of Sanov's Theorem

Let $\Pi_n = \Pi \cap \mathbb{P}_n$ be the set of possible n-types in Π, then

- $\text{Prob}\{\hat{P}_n \in \Pi_n\} = P^n(\cup_{Q\in\Pi_n} \mathcal{T}_Q^n) = \sum_{Q\in\Pi_n} P^n(\mathcal{T}_Q^n)$ and
- $\sum\limits_{Q\in\Pi_n} P^n(\mathcal{T}_Q^n) \leq \sum\limits_{Q\in\Pi_n} 2^{-nD(\Pi_n||P)} \leq \binom{n+M-1}{M-1} 2^{-nD(\Pi_n||P)}$
  since $P^n(\mathcal{T}_Q^n) \leq 2^{-nD(Q||P)} \leq 2^{-nD(\Pi_n||P)}$ and
- $\sum\limits_{Q\in\Pi_n} P^n(\mathcal{T}_Q^n) \geq \sum\limits_{Q\in\Pi_n} \frac{1}{\binom{n+M-1}{M-1}} 2^{-nD(Q||P)} \geq \frac{1}{\binom{n+M-1}{M-1}} 2^{-nD(\Pi_n||P)}$

Result follows taking the limit of $-\frac{1}{n}\log$ of the RHS and LHS. $\square$

Pythagorean theorem

- $D(P||Q)$ is not a metric, but it behaves like an Euclidean metric

Theorem 1: For a closed convex set of distributions $\mathcal{E} \subset \mathcal{P}$ and distribution $Q \notin \mathcal{E}$.

$$D(P^*||Q) = \min_{P \in \mathcal{E}} D(P||Q)$$

then

$$D(P||Q) \geq D(P||P^*) + D(P^*||Q) \quad \forall P \in \mathcal{E}.$$

- The result implies if you have a sequence $P_n \in \mathcal{E}$ with $D(P_n||Q) \overset{n\to\infty}{\longrightarrow} D(P^*||Q)$, then $D(P_n||P^*) \overset{n\to\infty}{\longrightarrow} 0$.

Proof of Pythagorean theorem

- Let $P_\lambda = \lambda P + (1-\lambda)P^* \in \mathcal{E}$, since $\mathcal{E}$ is convex, $P_\lambda \xrightarrow{\lambda \to 0} P^*$.
- Since $D(P^*||Q) \leq D(P_\lambda||Q) = D_\lambda$, we have $\frac{\mathrm{d}\, D_\lambda}{\mathrm{d}\, \lambda}|_{\lambda=0} \leq 0$.

$$\frac{\mathrm{d}\, D_\lambda}{\mathrm{d}\, \lambda} = \frac{\mathrm{d}}{\mathrm{d}\, \lambda}\Big[\sum P_\lambda(x) \log \frac{P_\lambda(x)}{Q(x)}\Big] = \sum (P(x) - P^*(x)) \log \frac{P_\lambda(x)}{Q(x)}$$

since $\sum_x P(x) - P^*(x) = 0$. For $\lambda = 0$ we have $P_\lambda = P^*$ and

$$\begin{aligned} 0 \leq \frac{\mathrm{d}\, D_\lambda}{\mathrm{d}\, \lambda}\Big|_{\lambda=0} &= \sum (P(x) - P^*(x)) \log \frac{P^*(x)}{Q(x)} \\ &= \sum P(x) \log \frac{P^*(x)}{Q(x)} \frac{P(x)}{P(x)} - \sum P^*(x) \log \frac{P^*(x)}{Q(x)} \\ &= D(P||Q) - D(P||P^*) - D(P^*||Q) \end{aligned}$$

$\square$

Conditional Limit Theorem

Consider a set of distributions $\mathcal{E}$, e.g. satisfying a condition.

- Sanov: For sequence generated by distribution $Q \notin \mathcal{E}$, the probability that the sequence has a type in $\mathcal{E}$ is asymptotically dominated by the distribution in $\mathcal{E}$ that is closest to Q.
- The next theorem states that conditional probability of each random variable in the sequences asymptotically in probability also behaves as the dominating distribution.

Theorem 2: Let $\mathcal{E} \subset \mathcal{P}$ be a closed and convex set of distributions on $\mathcal{A}$ and $Q \notin \mathcal{E}$ a distribution on $\mathcal{A}$. Let sequence $x^n \in \mathcal{A}^n$ be a realization of independently drawn random variables $X_i \sim Q$ and P^* achieve $\min_{P \in \mathcal{E}} D(P||Q) = D(P^*||Q)$. Then

$$\mathrm{Prob}\{X_1 = a | \hat{P}_{X^n} \in \mathcal{E}\} \overset{n\to\infty}{\longrightarrow} P^*(a)$$

in probability (with respect to X^n).[2]

[2] Convergence in probability: $\lim_{n\to\infty} \mathrm{Prob}\{|Z_n - Z| \geq \epsilon\} = 0$ for $\epsilon > 0$.

Proof of Conditional Limit Theorem

- Let $D^* = D(P^*||Q) = \min_{P\in\mathcal{E}} D(P||Q)$ with P^* unique since $D(P||Q)$ is strictly convex in P and convex set $\mathcal{S}_t = \{P \in \mathcal{P} : D(P||Q) \leq t\}$. Therewith define

$$\mathcal{U}_1 = \mathcal{S}_{D^*+\delta} \cap \mathcal{E} \qquad \mathcal{U}_2 = \mathcal{S}_{D^*+2\delta} \cap \mathcal{E} \qquad \mathcal{V} = \mathcal{E} \setminus \mathcal{U}_2.$$

- For $P \in \mathcal{V}$ we have $Q^n(\mathcal{T}_P^n) \leq 2^{-nD(P||Q)} \leq 2^{-n(D^*+2\delta)}$ and $(n+1)^M Q^n(\mathcal{T}_P^n) \geq 2^{-nD(P||Q)} \geq 2^{-n(D^*+\delta)}$ for $P \in \mathcal{U}_1$.

$$\mathrm{Prob}\{\hat{P}_{X^n} \in \mathcal{V} | \hat{P}_{X^n} \in \mathcal{E}\} = \frac{Q^n(\mathcal{V}\cap\mathcal{E})}{Q^n(\mathcal{E})} \leq \frac{Q^n(\mathcal{V})}{Q^n(\mathcal{U}_1)} = \frac{\sum_{P\in\mathcal{V}} Q^n(\mathcal{T}_P^n)}{\sum_{P\in\mathcal{U}_1} Q^n(\mathcal{T}_P^n)}$$

$$\geq \frac{\sum\limits_{P\in\mathcal{V}} 2^{-n(D^*+2\delta)}}{\sum\limits_{P\in\mathcal{U}_1} \frac{2^{-n(D^*+\delta)}}{(n+1)^M}} \geq \frac{(n+1)^M 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^M} 2^{-n(D^*+\delta)}} = \underbrace{(n+1)^{2M} 2^{-n(D^*+\delta)}}_{\overset{n\to\infty}{\longrightarrow} 0}$$

$$\Rightarrow \mathrm{Prob}\{\hat{P}_{X^n} \in \mathcal{U}_2 | \hat{P}_{X^n} \in \mathcal{E}\} \to 1 \text{ as } n \to \infty.$$

- For all $P \in \mathcal{U}_2$ we have $D(P||Q) \leq D^* + 2\delta$ so that

$$0 \leq D(P||P^*) + D(P^*||Q) \overset{\text{Pythagorean}}{\leq} D(P||Q) \leq D^* + 2\delta$$

Since $D(P^*||Q) = D^*$ we have $D(P||P^*) \leq 2\delta$.

- Since this all holds as well for $\hat{P}_{x^n} \in \mathcal{U}_2$ we have for $n \to \infty$

$$\mathrm{Prob}\{D(\hat{P}_{X^n}||P^*) \leq 2\delta|\hat{P}_{X^n} \in \mathcal{E}\} = \mathrm{Prob}\{\hat{P}_{X^n} \in \mathcal{U}_2|\hat{P}_{X^n} \in \mathcal{E}\} \to 1$$

- A small relative entropy implies a small L_1-distance[3] which implies a small $\max_{a \in \mathcal{A}} |\hat{P}_{X^n}(a) - P^*(a)|$ so that we have

$$\mathrm{Prob}\{|\hat{P}_{X^n}(a) - P^*(a)| \geq \epsilon|\hat{P}_{X^n} \in \mathcal{E}\} \overset{n \to \infty}{\longrightarrow} 0 \qquad \forall a \in \mathcal{A},$$

alternatively we can write $\mathrm{Prob}\{X_1 = a|\hat{P}_{X^n} \in \mathcal{E}\} \to P^*(a)$ as $n \to \infty$ in probability for all $a \in \mathcal{A}$. $\square$

[3] $D(P_1||P_2) \geq \frac{1}{2\log 2}||P_1 - P_2||_1^2$ see Lemma 11.6.1 [CT].

Hypothesis Testing

- Observation of n independent drawings x_i of random variable X_i with an **unknown** distribution Q on $\mathcal{A}$, $i = 1, \ldots, n$.
- Decision maker needs to decide between hypotheses

  $H_0 :\ Q = P_0$
  $H_1 :\ Q = P_1$

- Let $g : \mathcal{A}^n \to \{H_0, H_1\}$ denote a (non-)randomized test for sample size n characterized by **decision region** $\mathcal{D} \subseteq \mathcal{A}^n$:

$$g(x^n) = \begin{cases} H_0, & \text{if } x^n \in \mathcal{D}, \\ H_1, & \text{if } x^n \notin \mathcal{D}, \end{cases}$$

- Error terminology
  - **Type 1** error: $g(x^n) = H_1$, i.e., $x^n \notin \mathcal{D}$ although $Q = P_0$
    - Type 1 error probability: $\alpha = P_0^n(\mathcal{D}^c)$ with $\mathcal{D}^c = \mathcal{A}^n \setminus \mathcal{D}$.
  - **Type 2** error: $g(x^n) = H_0$, i.e., $x^n \in \mathcal{D}$ although $Q = P_1$
    - Type 2 error probability $\beta = P_1^n(\mathcal{D})$.

Neyman-Pearson Lemma

- Wish to find a test g that minimizes both probabilities of error α and β, but there is a trade-off.
- Neyman-Pearson approach is the constraint optimization problem:

$$\min_{\mathcal{D}\subseteq A^n} P_1^n(\mathcal{D}) \quad \text{subject to} \quad P_0^n(\mathcal{D}^c) \leq \epsilon$$

  - Ratio tests $\frac{P_0^n(x^n)}{P_1^n(x^n)} \underset{H_1}{\overset{H_0}{\gtrless}} T$ will be sufficient for optimality since...

Neyman-Pearson Lemma: Let $X_i \overset{iid}{\sim} Q$ defined on finite set $\mathcal{A}$, $i = 1, \ldots n$. Consider the decision problem with hypothesis $H_0 : Q = P_0$ and $H_1 : Q = P_1$. For $T \geq 0$ define decision region

$$\mathcal{D}_n(T) = \left\{ x^n \in \mathcal{A}^n \, : \, \frac{P_0^n(x^n)}{P_1^n(x^n)} > T \right\}$$

with associated error probabilities $\alpha^* = P_0^n(\mathcal{D}_n^c(T))$ and $\beta^* = P_1^n(\mathcal{D}_n(T))$. Let $\mathcal{F}$ be any other decision region with associated error probabilities α and β. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Proof Neyman-Pearson Lemma

- Let $\mathcal{D} = \mathcal{D}_n(T)$ and let $\mathcal{F}$ denote any other decision region. Let $\mathbb{1}_{\mathcal{D}}$ and $\mathbb{1}_{\mathcal{F}}$ denote corresponding indicator functions
- For any $x^n \in \mathcal{A}^n$ we have[4]

$$\big(\mathbb{1}_{\mathcal{D}}(x^n) - \mathbb{1}_{\mathcal{F}}(x^n)\big)\big(P_0(x^n)) - T \cdot P_1(x^n)\big) \geq 0$$

- Summing over all $x^n \in \mathcal{A}^n$ and expanding the product gives

$$\begin{aligned} 0 &\leq \sum \big(\mathbb{1}_{\mathcal{D}} P_0 - T\mathbb{1}_{\mathcal{D}} P_1 - \mathbb{1}_{\mathcal{F}} P_0 + T\mathbb{1}_{\mathcal{F}} P_1\big) \\ &= \underbrace{\sum_{x^n \in \mathcal{D}} (P_0 - TP_1)}_{=(1-\alpha^*)-T\beta^*} - \underbrace{\sum_{x^n \in \mathcal{F}} (P_0 - TP_1)}_{=(1-\alpha)-T\beta} = T(\beta - \beta^*) - (\alpha^* - \alpha) \end{aligned}$$

since $T \geq 0$ it follows that if $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$. $\square$

[4] If $x^n \in \mathcal{D}$ both factors are ≥ 0 and if $x^n \notin \mathcal{D}$, then both factors are ≤ 0.

- Q: What to expect if support$(P_0) \cap$ support$(P_1) \neq \emptyset$?

Theorem 3: Let P_0 and P_1 be any two distributions on $\mathcal{A}$ and suppose a sequence of sets $\mathcal{B}_n \subseteq \mathcal{A}^n$ that satisfies $P_0^n(\mathcal{B}_n) \geq \gamma$ for all n and a given positive $\gamma > 0$.[5] Then

$$\liminf_{n\to\infty} \frac{1}{n} \log P_1^n(\mathcal{B}_n) \geq -D(P_0||P_1).$$

Proof: Let $\delta_n = \frac{|\mathcal{A}|\log n}{n}$. Then $2^{-n\delta_n} = n^{-M}$ so that we have $\binom{n+M-1}{M-1} 2^{-n\delta_n} \leq \frac{(n+1)^{M-1}}{n^M} \stackrel{n\to\infty}{\longrightarrow} 0$. For a sample x^n drawn $\stackrel{iid}{\sim} P_0$, from a previous corollary we have $\mathrm{Prob}\{D(\hat{P}_{X^n}||P_0) \geq \delta_n\} \leq \binom{n+M-1}{M-1} 2^{-n\delta_n} \stackrel{n\to\infty}{\longrightarrow} 0$. Thus,

$$\mathrm{Prob}\{D(\hat{P}_{X^n}||P_0) < \delta_n\} = \sum_{Q:D(Q||P_0)<\delta_n} P_0^n(\mathcal{T}_Q^n) \stackrel{n\to\infty}{\longrightarrow} 1.$$

---

[5] If $\mathcal{B}^n$ is the decision region, then type 1 error is non-trivially bounded $P_0^n(\mathcal{B}_n^c) = 1 - P_0^n(\mathcal{B}_n) \leq 1 - \gamma < 1\, \forall n$.

- From the assumption $P_0^n(\mathcal{B}_n) \geq \gamma$ for all n it follows

$$\exists n_0 : \sum_{Q: D(Q||P_0) < \delta_n} P_0^n(\mathcal{T}_Q^n \cap \mathcal{B}_n) > \frac{\gamma}{2} \quad \forall n > n_0.$$

- Consequently, there exists n-types Q_n with $D(Q_n||P_0) < \delta_n$ and $P_0^n(\mathcal{T}_{Q_n}^n \cap \mathcal{B}_n) \geq \frac{\gamma}{2} P_0^n(\mathcal{T}_{Q_n}^n)$ for all $n > n_0$.
- Since sequences of the same type are equiprobable, which holds for any distribution P on $\mathcal{A}$, the last inequality holds also for P_1. Thus, for $n > n_0$ we have

$$P_1^n(\mathcal{B}_n) \geq P_1^n(\mathcal{T}_{Q_n}^n \cap \mathcal{B}_n) \geq \frac{\gamma}{2} P_1^n(\mathcal{T}_{Q_n}^n) \geq \frac{\gamma}{2} \frac{1}{\binom{n+M-1}{M-1}} 2^{-nD(Q_n||P_1)}$$

- $D(Q_n||P_0) < \delta_n \to 0$ implies $D(Q_n||P_1) \overset{n\to\infty}{\longrightarrow} D(P_0||P_1)$

$$\frac{1}{n} \log P_1^n(\mathcal{B}_n) \geq \underbrace{-\frac{1}{n} \log\left[\frac{2}{\gamma} \binom{n+M-1}{M-1}\right]}_{\overset{n\to\infty}{\longrightarrow} 0} + \underbrace{D(Q_n||P_1)}_{\overset{n\to\infty}{\longrightarrow} D(P_0||P_1)} \qquad \square$$

Testing null-hypothesis formulation

- Observation of n independent drawings from an **unknown** distribution P on $\mathcal{A}$ denoted by x^n.
- Testing of *null-hypothesis*: unknown P belongs to a given set of distributions Π on $\mathcal{A}$
- (Non-)randomized test for samples size n is characterized by **critical region** $\mathcal{C} \subseteq \mathcal{A}^n$:
  - null-hypothesis is accepted if $x^n \notin \mathcal{C}$ and rejected otherwise
- Error terminology
  - **Type 1** error: Null-hypothesis rejected although $P \in \Pi$
    - Type 1 error probability is given by $P^n(\mathcal{C})$
  - **Type 2** error: Null-hypothesis accepted although $P \notin \Pi$
    - Type 2 error probability $P^n(\mathcal{C}^c)$ with $\mathcal{C}^c = \mathcal{A} \setminus \mathcal{C}$
- Since $P \in \Pi$ is unknown we now may require tests with desired performance for all $P \in \Pi$, e.g. bounded type 1 error $P^n(\mathcal{C}) \leq \epsilon$ for all $P \in \Pi$ and characterize the decaying type 2 error for all $P \notin \Pi$!

Theorem 4: Consider testing the null-hypothesis that $P \in \Pi$, where $\Pi \subset \mathcal{P}$ is a closed set of distributions on $\mathcal{A}$. Then tests with critical region

$$\mathcal{C}_n = \left\{ x^n \in \mathcal{A}^n : \inf_{P \in \Pi} D(\hat{P}_{x^n} || P) \geq \delta_n \right\} \quad \text{with } \delta_n = \frac{|\mathcal{A}| \log n}{n}$$

have type 1 error probability $P^n(\mathcal{C}_n)$ not exceeding ϵ_n, where $\epsilon_n \to 0$, and for each $Q \notin \Pi$, the type 2 error probability $Q^n(\mathcal{C}_n^c)$ goes to 0 with exponential rate $D(\Pi || Q)$.

- Considering the previous hypothesis testing problem deciding between distributions P_0 and P_1, the result above (with $\Pi = \{P_2\}$) shows the existence of sets $\mathcal{B}_n \subset \mathcal{A}^n$ satisfying

$$P_0^n(\mathcal{B}_n) \to 1 \qquad \frac{1}{n} \log P_1^n(\mathcal{B}_n) \to -D(P_1 || P_2)$$

as $n \to \infty$. This result is known as Stein's Lemma.[6]

[6] Stein's Lemma can be also proved using a weak typicality argument so that it applies to continuous distributions with finite relative entropy, see [CT].

Proof of theorem:

- For type 1 error, same arguments as proof of previous corollary

$$P^n(\mathcal{C}_n) = \sum_{Q:\, \inf_{P\in\Pi} D(Q||P) \geq \delta_n} \underbrace{P^n(\mathcal{T}_Q^n)}_{\leq 2^{-nD(Q||P)}} \leq \binom{n+M-1}{M-1} 2^{-n\delta_n} = \epsilon_n \overset{n\to\infty}{\longrightarrow} 0$$

- For type 2 error, for each $Q \notin \Pi$ we have

$$Q^n(\mathcal{C}_n^c) = \sum_{R:\, \inf_{P\in\Pi} D(R||P) < \delta_n} \underbrace{Q^n(\mathcal{T}_R^n)}_{\leq 2^{-nD(R||Q)}} \leq \binom{n+M-1}{M-1} 2^{-n\xi_n}$$

  with $\xi_n = \inf_{R:\, \inf_{P\in\Pi} D(R||P) < \delta_n} D(R||Q)$

- Since $\lim_{n\to\infty} \xi_n = \inf_{P\in\Pi} D(P||Q) = D(\Pi||Q)$ so that

$$\limsup_{n\to\infty} \frac{1}{n} \log Q^n(\mathcal{C}_n^c) \leq -D(\Pi|Q) \qquad \square$$

Combining results

- Theorem 3 can be applied using $\mathcal{C}_n^c$ defined in Theorem 4 as sets $\mathcal{B}_n$ as follows: For any $P \in \Pi$
  - we have $P^n(\mathcal{C}_n) \leq \epsilon_n < 1$ with $\epsilon_n \to 0$ for the type 1 error.

  ⇒ There exists $\delta > 0$ such that $\epsilon_n \leq 1 - \delta$ so that

$$P^n(\mathcal{C}_n^c) = 1 - P^n(\mathcal{C}_n) \geq 1 - \epsilon_n \geq \delta > 0$$

  - Thus, Theorem 3 can be applied for any $P_1 \notin \Pi$ so that

$$\liminf_{n\to\infty} \frac{1}{n} \log P_1^n(\mathcal{C}_n^c) \geq -D(\Pi||P_1) \quad \forall P_1 \notin \Pi$$

- The combination of the previous with Theorem 4 results in

$$\lim_{n\to\infty} \frac{1}{n} \log P_1^n(\mathcal{C}_n^c) = -D(\Pi||P_1) \quad \forall P_1 \notin \Pi$$

- Hence, the test related to $\mathcal{C}_n$ are asymptotically optimal.[7]
  - Closedness of Π in Theorem 4 ensures $D(\Pi||P_1) > 0$ if $P_1 \notin \Pi$, i.e. exponential decay rate for all P_2

[7] Criterion $\inf\limits_{P\in\Pi} D(\hat{P}_{x^n}||P) \geq \delta_n \Leftrightarrow \frac{\sup_{P\in\Pi} P^n(x^n)}{Q(x^n)} \leq 2^{-n\delta_n}$ with $Q = P_{x^n}$.

Bayesian setting – Chernonff information

- Consider the two hypothesis setting with **prior** probabilities.
  - $X_1, \ldots, X_n \overset{iid}{\sim} Q$ with hypotheses $H_0 : Q = P_0$ and $H_1 : Q = P_1$ with prior probabilities π_0 and π_1
  - Objective is probability of error $P_e^{(n)} = \pi_0 \alpha_n + \pi_1 \beta_n$ with

$$D^* = \lim_{n \to \infty} -\frac{1}{n} \log \min_{\mathcal{D}_n \subset \mathcal{A}^n} P_e^{(n)}$$

Theorem 5: (Chernoff) The best achievable exponent for the Bayesian probability of error is given by

$$D^* = D(P_{\lambda^*} || P_1) = D(P_{\lambda^*} || P_2)$$

with $P_\lambda(x) = \frac{P_0^\lambda(x) P_1^{1-\lambda}(x)}{\sum\limits_{a \in \mathcal{A}} P_1^\lambda(a) P_2^{1-\lambda}(a)}$ and λ^* the value of λ such that $D(P_{\lambda^*} || P_0) = D(P_{\lambda^*} || P_1)$.

- It can be shown that D^* is equivalent to the standard definition of Chernoff information $C(P_1, P_2) = -\min\limits_{0 \leq \lambda \leq 1} \log \left[\sum_{a \in \mathcal{A}} P_0^\lambda(a) P_1^{1-\lambda}(a) \right]$

Proof

- The Neyman-Pearson optimal test can be written as (HW):

$$D(\hat{P}_{x^n}||P_1) - D(\hat{P}_{x^n}||P_0) \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{n}\log T$$

- Let $\mathcal{D}_n$ denote the set of types associated with hypothesis H_0 and $\mathcal{D}_n^c$ is the set of types associated with hypothesis H_1, then we have $\alpha_n = P_0^n(\mathcal{D}_n^c)$ and $\beta_n = P_1^n(\mathcal{D}_n)$
- $\min_P D(P||P_1)$ subject to $D(P||P_0) - D(P||P_1) \geq \frac{1}{n}\log T$ provides type $\hat{P}_{x^n} \in \mathcal{D}_n$ closest to P_1 but still deciding for H_0
  - Simple calculus shows that P_λ is minimizer [CT (11.200)] where λ is chosen such that $D(P_\lambda||P_0) - D(P_\lambda||P_1) = \frac{1}{n}\log T$
- From Sanov's theorem we have
  - $-\frac{1}{n}\log\alpha_n = -\frac{1}{n}\log P_0^n(\mathcal{D}_n^c) \overset{n\to\infty}{\longrightarrow} D(\mathcal{D}_n^c||P_0) = D(P_\lambda||P_0)$
  - $-\frac{1}{n}\log\beta_n = -\frac{1}{n}\log P_1^n(\mathcal{D}_n) \overset{n\to\infty}{\longrightarrow} D(\mathcal{D}_n||P_1) = D(P_\lambda||P_1)$

$$\lim_{n\to\infty} -\frac{1}{n}\log P_e^{(n)} = \min\{D(P_\lambda||P_0), D(P_\lambda||P_1)\}$$

$\Rightarrow$ The optimal T is where $D(P_\lambda||P_0) = D(P_\lambda||P_1) \Rightarrow \lambda^*$. $\square$