EEE 551 Information Theory (Spring 2022)

Chapter 11: Information Theory and Statistics

The Method of Types

- Consider a finite alphabet $\mathcal{X}$
- For a sequence $x^n \in \mathcal{X}^n$ and $a \in \mathcal{X}$, let

$$\begin{aligned} N(a|x^n) &= \#\text{ of occurrences of } a \text{ in } x^n \\ &= |\{i : x_i = a\}| \\ &= \sum_{i=1}^{n} \mathbf{1}(x_i = a) \end{aligned}$$

- $P_{x^n}(a) = \dfrac{N(a|x^n)}{n}$ is called the **type** of x^n
- For example, if $x^n = (0, 1, 1, 0, 0, 1, 0)$, then

$$N(0|x^n) = 4, \quad N(1|x^n) = 3$$

$$P_{x^n}(0) = \frac{4}{7}, \quad P_{x^n}(1) = \frac{3}{7}$$

- The type is a distribution: $\displaystyle\sum_{a\in\mathcal{X}} P_{x^n}(a) = \sum_{a\in\mathcal{X}} \frac{N(a|x^n)}{n} = \frac{n}{n} = 1$

The Simplex and the Set of Types

- Let $\mathcal{P}$ be the **probability simplex** for $\mathcal{X}$, the set of probability distributions on $\mathcal{X}$:

$$\mathcal{P} = \left\{ P \in \mathbb{R}^{|\mathcal{X}|} : P(x) \geq 0 \text{ for all } x \in \mathcal{X},\ \sum_{x \in \mathcal{X}} P(x) = 1 \right\}$$

- Let $\mathcal{P}_n$ be the set of all types of n-length sequences
- For example, if $\mathcal{X} = \{0, 1\}$, then

$$\mathcal{P}_n = \left\{ (0,1), \left(\frac{1}{n}, \frac{n-1}{n}\right), \left(\frac{2}{n}, \frac{n-2}{n}\right), \ldots, \left(\frac{n-1}{n}, \frac{1}{n}\right), (1,0) \right\}$$

- $\mathcal{P}_n \subset \mathcal{P}$
- $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$

  Proof:

  - For any type $P \in \mathcal{P}_n$ and each $a \in \mathcal{X}$, $P(a) \in \left\{ 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1 \right\}$
  - Thus at most $n+1$ choices for each $P(a)$
  - Therefore $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$

  This bound is loose, but what matters is that the number of types is **polynomial** in n, whereas the number of sequences is **exponential**. Thus there are exponentially many sequences with each type

Probability of a Sequence

For distribution $Q(x) \in \mathcal{P}$, define i.i.d. distribution $Q^n(x^n) = \prod_{i=1}^{n} Q(x_i)$

$Q^n(x^n) = 2^{-n[D(P_{x^n} \| Q) + H(P_{x^n})]}$

Proof:

$$
\begin{aligned}
\log Q^n(x^n) &= \sum_{i=1}^{n} \log Q(x_i) \\
&= \sum_{a \in \mathcal{X}} N(a|x^n) \log Q(a) \\
&= \sum_{a \in \mathcal{X}} nP_{x^n}(a) \log Q(a) \\
&= n \sum_{a \in \mathcal{X}} P_{x^n}(a) \big[\log Q(a) - \log P_{x^n}(a) + \log P_{x^n}(a) \big] \\
&= n \sum_{a \in \mathcal{X}} P_{x^n}(a) \left[-\log \frac{P_{x^n}(a)}{Q(a)} + \log P_{x^n}(a) \right] \\
&= n \big[-D(P_{x^n} \| Q) - H(P_{x^n}) \big]
\end{aligned}
$$

Corollary: If $x^n \in T(Q)$, then $Q^n(x^n) = 2^{-nH(Q)}$

Type Class

Given a type P, the **type class** $T(P)$ is the set of n-length sequences with type P; i.e.

$$T(P) = \{x^n \in \mathcal{X}^n : P_{x^n} = P\}$$

Example: $\mathcal{X} = \{1, 2, 3\}$, $n = 5$, $P(1) = \frac{3}{5}$, $P(2) = \frac{1}{5}$, $P(3) = \frac{1}{5}$

$$\begin{aligned} T(P) = \{&11123, 11132, 11213, 11231, 11312, 11321, 12113, \\ &12131, 12311, 13112, 13121, 13211, 21113, 21131, \\ &21311, 23111, 31112, 31121, 31211, 32111\} \end{aligned}$$

$$|T(P)| = \frac{5!}{3!\,1!\,1!} = \binom{5}{3, 1, 1} = 20$$

Size of Type Class

For any type $P \in \mathcal{P}_n$, $|T(P)| = \dfrac{n!}{\prod_{x \in \mathcal{X}} (nP(x))!}$

We may more usefully bound the type class size as follows:

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}$$

Proof of upper bound: Let $P^n(T(P)) = \Pr\{X^n \in T(P)\}$ where $X^n \overset{\text{iid}}{\sim} P(x)$.

$$\begin{aligned} 1 &\ge P^n(T(P)) \\ &= \sum_{x^n \in T(P)} P^n(x^n) \\ &= \sum_{x^n \in T(P)} 2^{-nH(P)} \\ &= |T(P)| 2^{-nH(P)} \end{aligned}$$

Proof of lower bound:

We will prove that $P^n(T(P)) \geq P^n(T(Q))$ for all $Q \in \mathcal{P}_n$. Therefore:

$$\begin{aligned}
1 &= \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \\
&\leq \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \\
&= |\mathcal{P}_n| P^n(T(P)) \\
&\leq (n+1)^{|\mathcal{X}|} P^n(T(P)) \\
&= (n+1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)}
\end{aligned}$$

Rearranging gives $|T(P)| \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)}$

To prove $P^n(T(P)) \geq P^n(T(Q))$, we need the following fact:

For any integers m, k, $\quad \dfrac{m!}{k!} \geq k^{m-k}$

- If $m \geq k$, then $\dfrac{m!}{k!} = \prod_{i=k+1}^{m} i \geq k^{m-k}$
- If $m < k$, then $\dfrac{k!}{m!} = \prod_{i=m+1}^{k} i \leq k^{k-m}$, so $\dfrac{m!}{k!} \geq k^{m-k}$

Thus:

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(Q))} &= \frac{|T(P)| \ \prod_x P(x)^{nP(x)}}{|T(Q)| \ \prod_x P(x)^{nQ(x)}} \\
&= \frac{\frac{n!}{\prod_x (nP(x))!}}{\frac{n!}{\prod_x (nQ(x))!}} \prod_x P(x)^{n(P(x)-Q(x))} \\
&= \prod_x \frac{(nQ(x))!}{(nP(x))!} \, P(x)^{n(P(x)-Q(x))} \\
&\geq \prod_x (nP(x))^{nQ(x)-nP(x)} \, P(x)^{n(P(x)-Q(x))} \\
&= \prod_x n^{n(Q(x)-P(x))} \\
&= n^{n \sum_x (Q(x)-P(x))} = 1
\end{aligned}$$

Probability of Type Class

For any $P \in \mathcal{P}_n$ and any distribution Q,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)}$$

Proof:

$$Q^n(T(P)) = \sum_{x^n \in T(P)} Q^n(x^n) = |T(P)| 2^{-n[D(P\|Q)+H(P)]}$$

From upper bound on $|T(P)|$:

$$Q^n(T(P)) \leq 2^{-nD(P\|Q)}$$

From lower bound on $|T(P)|$:

$$Q^n(T(P)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)}$$

Summary of Results on the Method of Types

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$$

$$Q^n(x^n) = 2^{-n[D(P_{x^n}\|Q)+H(P_{x^n})]}$$

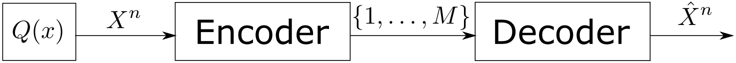
$$|T(P)| \doteq 2^{nH(P)}$$

$$Q^n(T(P)) \doteq 2^{-nD(P\|Q)}$$

where $\doteq$ means equality in first-order in the exponent

i.e. $a_n \doteq b_n$ iff $\quad \lim_{n\to\infty} \frac{1}{n} \log a_n = \lim_{n\to\infty} \frac{1}{n} \log b_n$

Universal Source Coding



Source distribution is i.i.d. but **unknown** — code must work no matter what Q is

An (M, n) code is given by

- An encoding function $f : \mathcal{X}^n \to \{1, \ldots, M\}$
- A decoding function $g : \{1, \ldots, M\} \to \mathcal{X}^n$

Probability of error with respect to distribution Q is

$$P_e^{(n)}(Q) = Q^n \{g(f(X^n)) \neq X^n)\}$$

Theorem

For any rate R, there exists a sequence of $(2^{nR}, n)$ codes such that

$$P_e^{(n)}(Q) \to 0 \text{ as } n \to \infty \text{ for all } Q \text{ such that } H(Q) < R.$$

Proof:

- Fix rate R. Let $R_n = R - |\mathcal{X}|\dfrac{\log(n+1)}{n}$
- Let $A = \{x^n \in \mathcal{X}^n : H(P_{x^n}) \leq R_n\}$
- Encoder: $f(x^n) = \begin{cases} \text{index of } x^n \in A, & \text{if } x^n \in A \\ 1, & \text{otherwise} \end{cases}$

  Decoder: given $f(x^n) = m$, select $\hat{x}^n \in A$ where $f(\hat{x}^n) = m$
- Note that $P_e^{(n)}(Q) = Q^n(A^c)$
- Need to show: (1) $|A| \leq 2^{nR}$, (2) For any Q with $H(Q) < R$, $Q^n(A^c) \to 0$
- Proof of (1):

$$\begin{aligned}
|A| &= \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \\
&\leq (n+1)^{|\mathcal{X}|} 2^{nR_n} \\
&= 2^{n\left(R_n + \frac{|\mathcal{X}|\log(n+1)}{n}\right)} = 2^{nR}
\end{aligned}$$

Proof of (2): Assume $H(Q) < R$:

$$\begin{aligned} Q^n(A^c) &= \sum_{P \in \mathcal{P}_n : H(P) > R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{P \in \mathcal{P}_n : H(P) > R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{P \in \mathcal{P}_n : H(P) > R_n} 2^{-nD(P\|Q)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P : H(P) > R_n} D(P\|Q)} \end{aligned}$$

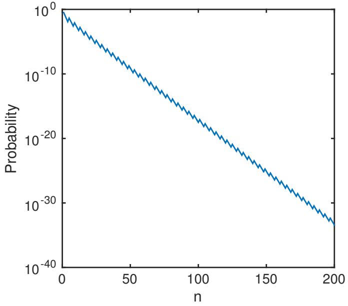
Since $R_n \to R$, for sufficiently large n, $H(Q) < R_n$

Thus $\min_{P : H(P) > R_n} D(P\|Q) > 0$, so $Q^n(A^c) \to 0$

Large Deviation Theory

Bounds on the probability that an i.i.d. sum differs significantly from its mean

Example: $X^n \overset{\text{iid}}{\sim}$ Bern$(1/3)$, how does $\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > 3/4\right\}$ behave for large n?



Probability roughly $2^{-nD^\star}$ for a constant $D^\star$

This event can be described in terms of the type P_{X^n}:

$$P_{X^n} \in E = \{P : P(1) > 3/4\}$$

Sanov's Theorem

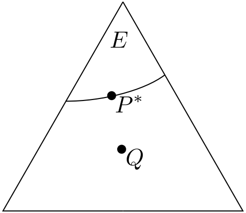
Theorem (Sanov's theorem)

Let $X^n \overset{iid}{\sim} Q(x)$, and let E be a set of probability distributions. Let

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

- $Q^n(E) = \Pr\{P_{X^n} \in E\} \le (n+1)^{|\mathcal{X}|} 2^{-nD(P^*\|Q)}$
- *If E is the closure of its interior*[1], *then*

$$\lim_{n\to\infty} \frac{1}{n} \log Q^n(E) = -D(P^*\|Q)$$



[1] Equivalent to the following: For all $a \in E$, there exists a sequence $a_1, a_2, \ldots$ where $a_n \to a$, and for each n, there exists $\epsilon_n > 0$ where $\{b : \|b - a_n\|_2 \le \epsilon_n\} \subset E$.

Proof of Sanov's theorem:

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\le \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P\|Q)} \\ &\le \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P\|Q)} \\ &\le (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in E} D(P\|Q)} \\ &= (n+1)^{|\mathcal{X}|} 2^{-nD(P^*\|Q)} \end{aligned}$$

If E is the closure of its interior, then there exists a sequence of distributions $P_n \in E \cap \mathcal{P}_n$ where $P_n \to P^*$.

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\ge Q^n(T(P_n)) \\ &\ge \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n\|Q)} \end{aligned}$$

Thus

$$\liminf_{n\to\infty} \frac{1}{n} \log Q^n(E) \ge \liminf_{n\to\infty} \left[-\frac{|\mathcal{X}| \log(n+1)}{n} - D(P_n\|Q) \right] = -D(P^*\|Q)$$

Example of Sanov's Theorem

- Let X_i be i.i.d. with $\mathbb{E}[X_i] = \mu$
- Consider a probability of the form $\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + \epsilon\right\}$
- Equivalent to $\Pr\{P_{X^n} \in E\}$ where $E = \left\{P : \sum_{a \in \mathcal{X}} P(a)\, a \geq \mu + \epsilon\right\}$
- By Sanov's theorem, $Q^n(E) \leq (n+1)^{|\mathcal{X}|} 2^{-nD^\star}$ where

$$D^\star = \min_{P : \sum_{a \in \mathcal{X}} P(a)\, a \geq \mu + \epsilon} D(P\|Q)$$

- To minimize over P, we form the Lagrangian

$$L(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)} + \lambda \left(\mu + \epsilon - \sum_x P(x)\, x\right) + \nu \left(\sum_x P(x) - 1\right)$$

- To solve for P, we need

$$0 = \frac{\partial L(P)}{\partial P(x)} = \log \frac{P(x)}{Q(x)} + \frac{1}{\ln 2} - \lambda\, x + \nu$$

- Thus

$$P(x) = Q(x) 2^{\lambda\, x - \nu - 1/\ln 2} = \frac{Q(x) 2^{\lambda x}}{\sum_{a \in \mathcal{X}} Q(a) 2^{\lambda a}}$$

where $\lambda \geq 0$ is chosen so that $\mathbb{E}_P[X] = \mu + \epsilon$

Alternative Proof of Large Deviation Bound

- Let $X^n \overset{\text{iid}}{\sim} Q(x)$
- We use the **Chernoff bounding** approach: For any $t > 0$,

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + \epsilon\right\} &= \Pr\left\{t\sum_{i=1}^{n} X_i \geq nt(\mu + \epsilon)\right\} \\
&= \Pr\left\{2^{t\sum_{i=1}^{n} X_i} \geq 2^{nt(\mu+\epsilon)}\right\} \\
&\leq \frac{\mathbb{E}\left[2^{t\sum_{i=1}^{n} X_i}\right]}{2^{nt(\mu+\epsilon)}} && \text{Markov's inequality} \\
&= 2^{-nt(\mu+\epsilon)}\mathbb{E}\left[\prod_{i=1}^{n} 2^{tX_i}\right] \\
&= 2^{-nt(\mu+\epsilon)}\left(\mathbb{E}[2^{tX}]\right)^n \\
&= 2^{-n\left(t(\mu+\epsilon) - \log \mathbb{E}[2^{tX}]\right)}
\end{aligned}$$

- Thus

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + \epsilon\right\} \leq \min_{t>0} 2^{-n\left(t(\mu+\epsilon) - \log \mathbb{E}[2^{tX}]\right)} = 2^{-n\left(\max_{t>0} t(\mu+\epsilon) - \log \mathbb{E}[2^{tX}]\right)}$$

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu+\epsilon\right\} \leq 2^{-nD^\star} \text{ where } D^\star = \max_{t>0}\ t(\mu+\epsilon) - \log \mathbb{E}[2^{tX}]$$

- The optimal t will satisfy

$$0 = \frac{d}{dt}\left(t(\mu+\epsilon) - \log\sum_x Q(x)2^{tx}\right) = \mu+\epsilon - \frac{\sum_x Q(x)x2^{tx}}{\sum_x Q(x)2^{tx}}$$

- Let $P(x) = \dfrac{Q(x)2^{tx}}{\sum_{a\in\mathcal{X}} Q(a)2^{ta}}$, so $\dfrac{\sum_x Q(x)x2^{tx}}{\sum_x Q(x)2^{tx}} = \sum_x P(x)x = \mathbb{E}_P[X]$
- Thus, the optimal t is where $\mathbb{E}_P[X] = \mu+\epsilon$, and so

$$\begin{aligned} D^\star &= t(\mu+\epsilon) - \log\sum_x Q(x)2^{tx} \\ &= t\,\mathbb{E}_P[X] - \log\sum_x Q(x)2^{tx} \\ &= \sum_x txP(x) - \log\sum_x Q(x)2^{tx} \\ &= \sum_x P(x)\log\frac{2^{tx}}{\sum_a Q(a)2^{ta}} \\ &= \sum_x P(x)\log\frac{P(x)}{Q(x)} = D(P\|Q) \end{aligned}$$

- This proves that $D^\star = \min_{P:\sum_a P(a)a\geq\mu+\epsilon} D(P\|Q)$

Hypothesis Testing

- Given a variable $X \in \mathcal{X}$, we wish to distinguish between two hypotheses:
  - $H_0 : \; X \sim P_0$
  - $H_1 : \; X \sim P_1$

- Problem: design a function (a test) $g : \mathcal{X} \to \{0, 1\}$ that accurately determines which hypothesis is in force.
  i.e. $g(X) = 0$ means "I guess H_0" and $g(X) = 1$ means "I guess H_1"

- It is equivalent to specify the acceptance region $A = \{x : g(x) = 1\}$

- Two probabilities of error:

$$\alpha = \Pr\{g(X) = 0 \,|\, H_1\} = P_1(A^c)$$
$$\beta = \Pr\{g(X) = 1 \,|\, H_0\} = P_0(A)$$

  We wish both to be small, but there is a trade-off

Neyman-Pearson Lemma

Lemma (Neyman-Pearson)

For $T > 0$, let $g^*(x)$ be a likelihood ratio test where $g^*(x) = 1$ iff

$$\frac{P_1(x)}{P_0(x)} > T.$$

Let α^*, β^* be the corresponding probabilities of error.

For any other test $g(x)$ with probabilities of error α, β, if $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Proof: Let A be the acceptance region for g^*, i.e. $A = \left\{ x : \frac{P_1(x)}{P_0(x)} > T \right\}$.

For all x,

$$\left[g^*(x) - g(x) \right] \left[P_1(x) - T\, P_0(x) \right] \geq 0.$$

Indeed, consider the two cases:

- $x \in A$: Thus $\frac{P_1(x)}{P_0(x)} > T$, i.e. $P_1(x) - T\, P_0(x) > 0$.

  Also $g^*(x) = 1$, so $g^*(x) - g(x) \geq 0$
- $x \notin A$: Thus $P_1(x) - T\, P_0(x) \leq 0$, and $g^*(x) = 0$, so $g^*(x) - g(x) \leq 0$

- We proved that for all x, $\left[g^*(x) - g(x)\right]\left[P_1(x) - T\,P_0(x)\right] \geq 0$.
- Thus,

$$\begin{aligned}
0 &\leq \sum_x \left[g^*(x) - g(x)\right]\left[P_1(x) - T\,P_0(x)\right] \\
&= \sum_x \left[g^*(x)\,P_1(x) - T\,g^*(x)\,P_0(x) - g(x)\,P_1(x) + T\,g(x)\,P_0(x)\right] \\
&= P_1(g^*(X) = 1) - T\,P_0(g^*(X) = 1) - P_1(g(X) = 1) + T\,P_0(g(X) = 1) \\
&= (1 - \alpha^*) - T\,\beta^* - (1 - \alpha) + T\,\beta \\
&= T(\beta - \beta^*) - (\alpha^* - \alpha).
\end{aligned}$$

- If $\alpha \geq \alpha^*$, then $0 \leq T(\beta - \beta^*)$
- Since $T > 0$, we have $\beta - \beta^* \geq 0$, i.e. $\beta \geq \beta^*$

Chernoff-Stein Lemma

- Consider the hypothesis testing problem between two i.i.d. distributions:
  - $H_0 : \; X^n \overset{\text{iid}}{\sim} P_0(x)$
  - $H_1 : \; X^n \overset{\text{iid}}{\sim} P_1(x)$

  where the problem is to design a test $g : \mathcal{X}^n \to \{0, 1\}$.
- Let $\alpha_n = P_1^n(g(X^n) = 0)$ and $\beta_n = P_0^n(g(X^n) = 1)$.
- For fixed $\epsilon \in (0, 1)$, let $\beta_n^\epsilon = \min_{g : \alpha_n \leq \epsilon} \beta_n$

Lemma (Chernoff-Stein)

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1 \| P_0).$$

Proof:

- By the Neyman-Pearson lemma, the optimal test has acceptance region

$$A = \left\{ \frac{P_1^n(x^n)}{P_0^n(x^n)} > T \right\}$$

- Let $\alpha_n(T) = P_1^n \left(\frac{P_1^n(X^n)}{P_0^n(X^n)} \leq T \right)$
- Let T_n^ϵ be the largest T such that $\alpha_n(T) \leq \epsilon$. Then $\beta_n^\epsilon = P_0^n \left(\frac{P_1^n(X^n)}{P_0^n(X^n)} > T_n^\epsilon \right)$.
- $\alpha_n(T) = P_1^n \left(\log \frac{P_1^n(X^n)}{P_0^n(X^n)} \leq \log T \right) = P_1^n \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_0(X_i)} \leq \frac{1}{n} \log T \right)$
- Variables $\log \frac{P_1(X_i)}{P_0(X_i)}$ are i.i.d. with mean (under P_1) $D(P_1 \| P_0)$, so by the law of large numbers, for any $\delta > 0$:
  - if $\frac{1}{n} \log T \geq D(P_1 \| P_0) + \delta$ then $\alpha_n(T) \to 1$
  - if $\frac{1}{n} \log T \leq D(P_1 \| P_0) - \delta$ then $\alpha_n(T) \to 0$

  Thus $\frac{1}{n} \log T_n^\epsilon \to D(P_1 \| P_0)$

$$
\begin{aligned}
\beta_n^\epsilon &= P_0^n\left(\frac{P_1^n(X^n)}{P_0^n(X^n)} > T_n^\epsilon\right) \\
&= P_0^n\left(\frac{1}{n}\sum_{i=1}^n \log\frac{P_1(X_i)}{P_0(X_i)} > \frac{1}{n}\log T_n^\epsilon\right) \\
&= P_0^n\left(\sum_x P_{X^n}(x)\log\frac{P_1(x)}{P_0(x)} > \frac{1}{n}\log T_n^\epsilon\right) \\
&= P_0^n\left(D(P_{X^n}\|P_0) - D(P_{X^n}\|P_1) > \frac{1}{n}\log T_n^\epsilon\right)
\end{aligned}
$$

Since $\lim\limits_{n\to\infty}\frac{1}{n}\log T_n^\epsilon = D(P_1\|P_0)$,

$$
\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\log\beta_n^\epsilon &= \lim_{n\to\infty}\frac{1}{n}\log P_0^n\Big(D(P_{X^n}\|P_0) - D(P_{X^n}\|P_1) \geq D(P_1\|P_0)\Big) \\
&= -\min_{P:D(P\|P_0)-D(P\|P_1)\geq D(P_1\|P_0)} D(P\|P_0) \\
&\leq -\min_{P:D(P\|P_0)-D(P\|P_1)\geq D(P_1\|P_0)} \big[D(P_1\|P_0) + D(P\|P_1)\big] \\
&\leq -D(P_1\|P_0)
\end{aligned}
$$

with equality if $P = P_1$

Chernoff Information

Consider the **Bayesian** hypothesis testing problem, with two hypotheses:

- $H_0 : \; X^n \overset{\text{iid}}{\sim} P_0$, occurs with prior probability π_0
- $H_1 : \; X^n \overset{\text{iid}}{\sim} P_1$, occurs with prior probability π_1

where $\pi_0 + \pi_1 = 1$.

Given a test $g : \mathcal{X}^n \to \{0, 1\}$, the probability of error is given by

$$P_e^{(n)} = \pi_1 \alpha_n + \pi_0 \beta_n = \pi_1 P_1^n(g(X^n) = 0) + \pi_0 P_0^n(g(X^n) = 1).$$

Let $D^* = \lim_{n\to\infty} -\frac{1}{n} \log \min_g P_e^{(n)}$

Theorem

$D^* = D(P_{\lambda^*} \| P_1) = D(P_{\lambda^*} \| P_0)$ *where*

$$P_\lambda(x) = \frac{P_1(x)^\lambda P_0(x)^{1-\lambda}}{\sum_{a\in\mathcal{X}} P_1(a)^\lambda P_0(a)^{1-\lambda}}$$

and $\lambda^* \in [0, 1]$ *is such that* $D(P_{\lambda^*} \| P_1) = D(P_{\lambda^*} \| P_0)$. *This quantity is called the* ***Chernoff information****.*

Proof:

- By the Neyman-Pearson lemma, the optimal test will be a likelihood ratio test with acceptance region

$$A = \left\{ x^n : \frac{P_1^n(x^n)}{P_0^n(x^n)} > T \right\} = \left\{ x^n : D(P_{x^n}\|P_0) - D(P_{x^n}\|P_1) > \frac{1}{n}\log T \right\}$$

- Thus

$$\begin{aligned}\lim_{n\to\infty} -\frac{1}{n}\log\beta_n &= \lim_{n\to\infty} -\frac{1}{n}\log P_0^n(A) \\ &= \min_{P: D(P\|P_0)-D(P\|P_1)\geq \frac{1}{n}\log T} D(P\|P_0)\end{aligned}$$

- To solve this optimization, consider the Lagrangian

$$\sum_x P(x)\log\frac{P(x)}{P_0(x)} - \lambda\sum_x P(x)\log\frac{P_1(x)}{P_0(x)} + \nu\sum_x P(x)$$

- Differentiating with respect to $P(x)$:

$$\log\frac{P(x)}{P_0(x)} + 1 - \lambda\log\frac{P_1(x)}{P_0(x)} + \nu = 0$$

$$\log \frac{P(x)}{P_0(x)} + 1 - \lambda \log \frac{P_1(x)}{P_0(x)} + \nu = 0$$

- Rearranging gives $P(x) = \dfrac{P_1(x)^\lambda P_0(x)^{1-\lambda}}{2^{\nu'}} = P_\lambda(x)$
- Thus $\beta_n \doteq 2^{-nD(P_\lambda \| P_0)}$ where λ is chosen so that $D(P_\lambda \| P_0) - D(P_\lambda \| P_1) = \dfrac{1}{n} \log T$
- By a similar analysis, $\alpha_n \doteq 2^{-nD(P_\lambda \| P_1)}$ where again $D(P_\lambda \| P_0) - D(P_\lambda \| P_1) = \dfrac{1}{n} \log T$
- $$\begin{aligned} P_e^{(n)} &= \pi_1 \alpha_n + \pi_0 \beta_n \\ &\doteq \pi_1 2^{-nD(P_\lambda \| P_1)} + \pi_0 2^{-nD(P_\lambda \| P_0)} \\ &\doteq 2^{-n \min\{D(P_\lambda \| P_1), D(P_\lambda \| P_0)\}} \end{aligned}$$
- $\min\{D(P_\lambda \| P_1), D(P_\lambda \| P_0)\}$ is maximized when $D(P_\lambda \| P_1) = D(P_\lambda \| P_0)$, i.e. $\lambda = \lambda^*$
- Therefore $P_e^{(n)} \doteq 2^{-nD(P_{\lambda^*} \| P_1)} = 2^{-nD(P_{\lambda^*} \| P_0)}$

Parameter Estimation

- Let $\theta \in \Theta$ be an unknown parameter to be estimated from data X related to θ
- For each θ, there is a PDF $f(x;\theta)$ for the distribution of X given θ
- An **estimator** is a function $T : \mathcal{X} \to \Theta$ that produces an estimate $T(X)$ that should be close to θ
- **Example:** $X \sim \mathcal{N}(\theta, 1)$. An estimator is $T(X) = X$
- **Example:** $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta,)$. An estimator is $T(X_1, \ldots, X_n) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$
- The **bias** of an estimator T is $\mathbb{E}_\theta[T(X)] - \theta$
- An estimator is said to be **unbiased** if its bias is 0 for all θ; i.e., if

$$\mathbb{E}_\theta[T(X)] = \theta \text{ for all } \theta$$

- **Question:** How small can we make the mean-square error of an estimator? i.e.,

$$\mathbb{E}_\theta\left[(T(X) - \theta)^2\right]$$

Cramér-Rao Bound

Theorem

For any unbiased estimator $T(X)$ of the parameter θ,

$$\mathbb{E}_\theta[(T(X)-\theta)^2] \geq \frac{1}{J(\theta)}$$

where $J(\theta)$ is the **Fisher information**, *defined by*

$$J(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2$$

Proof:

- Consider the variable inside the expectation (sometimes called the **score**):

$$V = \frac{\partial}{\partial\theta}\ln f(X;\theta) = \frac{\frac{\partial}{\partial\theta} f(X;\theta)}{f(X;\theta)}$$

- $J(\theta) = \mathbb{E}_\theta[V^2]$
- The expectation of the score is

$$\begin{aligned}\mathbb{E}_\theta[V] &= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\, f(x;\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

- By the Cauchy-Schwartz inequality,

$$\Big(\mathbb{E}_\theta[(V - \mathbb{E}_\theta V)(T - \mathbb{E}_\theta T)]\Big)^2 \leq \mathbb{E}_\theta (V - \mathbb{E}_\theta V)^2 \, \mathbb{E}_\theta (T - \mathbb{E}_\theta T)^2$$

- We know $\mathbb{E}_\theta V = 0$, and by the assumption that T is unbiased, $\mathbb{E}_\theta T = \theta$
- The left-hand side of the above inequality becomes

$$\Big(\mathbb{E}_\theta[V(T - \theta)]\Big)^2 = \Big(\mathbb{E}_\theta[VT] - \mathbb{E}_\theta[V\theta]\Big)^2 = \Big(\mathbb{E}_\theta[VT]\Big)^2$$

- The right-hand side becomes

$$\mathbb{E}_\theta V^2 \, \mathbb{E}_\theta[(T - \theta)^2] = J(\theta) \, \mathbb{E}_\theta[(T - \theta)^2]$$

- So $\Big(\mathbb{E}_\theta[VT]\Big)^2 \leq J(\theta) \, \mathbb{E}_\theta[(T - \theta)^2]$
- Rearranging gives

$$\mathbb{E}_\theta[(T - \theta)^2] \geq \frac{\Big(\mathbb{E}_\theta[VT]\Big)^2}{J(\theta)}$$

$$
\begin{aligned}
\mathbb{E}_\theta[VT] &= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\, T(x)\, f(x;\theta)\, dx \\
&= \int \frac{\partial}{\partial\theta} f(x;\theta)\, T(x)\, dx \\
&= \frac{\partial}{\partial\theta} \int f(x;\theta), T(x)\, dx \\
&= \frac{\partial}{\partial\theta} \mathbb{E}_\theta T(X) \\
&= \frac{\partial}{\partial\theta}\, \theta \\
&= 1
\end{aligned}
$$

Therefore

$$
\mathbb{E}_\theta[(T(X) - \theta)^2] \geq \frac{1}{J(\theta)}
$$

Cramér-Rao Bound for i.i.d. Data

- Suppose, for each θ, we observe $X_1, X_2, \ldots, X_n$ i.i.d., that is

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- The score variable is

$$V = \frac{\partial}{\partial \theta} \ln f(X_1, \ldots, X_n; \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(X_i; \theta) = \sum_{i=1}^{n} V_i$$

where $V_i = \frac{\partial}{\partial \theta} \ln f(X_i; \theta)$

- The Fisher information for n-samples is

$$J_n(\theta) = \mathbb{E}_\theta V^2 = \mathbb{E}_\theta \left(\sum_{i=1}^{n} V_i \right)^2 = \sum_{i=1}^{n} \mathbb{E}_\theta V_i^2 = nJ(\theta)$$

- Now the Cramér-Rao bound says that for any unbiased T,

$$\mathbb{E}_\theta[(T(X_1, \ldots, X_n) - \theta)^2] \geq \frac{1}{nJ(\theta)}$$

- That is, in the best case the mean squared error for n samples goes down like $1/n$

Relationship Between Fisher Information and Differential Entropy

- Assume that the parametric PDF has the form $f(x;\theta) = f(x-\theta)$; i.e., θ shifts the distribution of X
- The Fisher information becomes

$$\begin{aligned} J(\theta) &= \int f(x-\theta) \left[\frac{\partial}{\partial \theta} \ln f(x-\theta) \right]^2 dx \\ &= \int f(x-\theta) \left[\frac{\partial}{\partial x} \ln f(x-\theta) \right]^2 dx \\ &= \int f(x) \left[\frac{\partial}{\partial x} \ln f(x) \right]^2 dx \end{aligned}$$

- Since in this case J does not depend on θ, we write this as $J(X)$

Theorem (de Bruijn's identity)

Let X have finite variance with PDF $f(x)$. Let $Z \sim \mathcal{N}(0,1)$ independent of X. Then

$$\left. \frac{\partial}{\partial t} h(X + \sqrt{t}\, Z) \right|_{t=0} = \frac{1}{2 \ln 2} J(X).$$