Elements of Information Theory

George Voutsadakis[1]

[1] Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1. Information Theory and Statistics
   - Method of Types
   - Law of Large Numbers
   - Universal Source Coding
   - Large Deviation Theory
   - Examples of Sanov's Theorem
   - Conditional Limit Theorem
   - Hypothesis Testing
   - Chernoff-Stein Lemma
   - Chernoff Information
   - Fisher Information and the Cramér-Rao Inequality

Subsection 1

Method of Types

Types

- Let $X_1, X_2, \ldots, X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$.
- We use the notation x^n and $\boldsymbol{x}$ interchangeably to denote a sequence $x_1, x_2, \ldots, x_n$.

Definition

The **type** $P_{\boldsymbol{x}}$ (or **empirical probability distribution**) of a sequence $x_1, x_2, \ldots, x_n$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$, i.e.,

$$P_{\boldsymbol{x}}(a) = \frac{N(a|\boldsymbol{x})}{n}, \quad \text{for all } a \in \mathcal{X},$$

where $N(a|\boldsymbol{x})$ is the number of times the symbol a occurs in the sequence $\boldsymbol{x} \in \mathcal{X}^n$.

- The type of a sequence $\boldsymbol{x}$ is denoted as $P_{\boldsymbol{x}}$.
- The type $P_{\boldsymbol{x}}$ is a probability mass function on $\mathcal{X}$.

The Probability Simplex

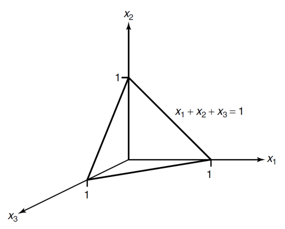
Definition

The **probability simplex** in $\mathbb{R}^m$ is the set of points

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m, \text{ such that } x_i \geq 0, \sum_{i=1}^{m} x_i = 1.$$

- The probability simplex is an $(m-1)$-dimensional manifold in m-dimensional space.
- When $m = 3$, the probability simplex is the set of points

$$\{(x_1, x_2, x_3) : \quad x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, \\ x_1 + x_2 + x_3 = 1\}.$$



Since this is a triangular two-dimensional flat in $\mathbb{R}^3$, we use a triangle to represent the probability simplex in later sections.

Type Classes

Definition

Let $\mathcal{P}_n$ denote the set of types with denominator n.

Example: If $\mathcal{X} = \{0,1\}$, the set of possible types with denominator n is

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

Definition

If $P \in \mathcal{P}_n$, the set of sequences of length n and type P is called the **type class** of P, denoted $T(P)$:

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

The type class is sometimes called the **composition class** of P.

Example

- Let $\mathcal{X} = \{1, 2, 3\}$, a ternary alphabet.

  Let $\boldsymbol{x} = 11321$.

  Then the type $P_{\boldsymbol{x}}$ is

$$P_{\boldsymbol{x}}(1) = \frac{3}{5}, \quad P_{\boldsymbol{x}}(2) = \frac{1}{5}, \quad P_{\boldsymbol{x}}(3) = \frac{1}{5}.$$

  The type class of $P_{\boldsymbol{x}}$ is the set of all sequences of length 5 with three 1's, one 2, and one 3,

$$T(P_{\boldsymbol{x}}) = \{11123, 11132, 11213, \ldots, 32111\}.$$

  The number of elements in $T(P)$ is

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

The Number of Types

Theorem

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

- There are $|\mathcal{X}|$ components in the vector that specifies $P_{\boldsymbol{x}}$.

  The numerator in each component can take on only $n+1$ values.

  So there are at most $(n+1)^{|\mathcal{X}|}$ choices for the type vector.
- Of course, these choices are not independent, e.g., the last choice is fixed by the others.
- The crucial point is that this is a polynomial as a function of n.
- The number of sequences is exponential in n.
- So at least one type has exponentially many sequences in its type class.
- In fact, the largest type class has essentially the same number of elements as the entire set of sequences, to first order in the exponent.

Product Distribution of i.i.d. Sequence

- Assume that the sequence $X_1, X_2, \ldots, X_n$ is drawn i.i.d. according to a distribution $Q(x)$.
- Let
$$Q^n(x^n) = \prod_{i=1}^n Q(x_i)$$
denote the product distribution associated with Q.

Theorem

If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. according to $Q(x)$, the probability of $\boldsymbol{x}$ depends only on its type and is given by

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

Product Distribution of i.i.d. Sequence (Cont'd)

- We compute

$$\begin{array}{rcl}
Q^n(\boldsymbol{x}) & = & \prod_{i=1}^n Q(x_i) \\
& = & \prod_{a\in\mathcal{X}} Q(a)^{N(a|\boldsymbol{x})} \\
& = & \prod_{a\in\mathcal{X}} Q(a)^{nP_{\boldsymbol{x}}(a)} \\
& = & \prod_{a\in\mathcal{X}} 2^{nP_{\boldsymbol{x}}(a)\log Q(a)} \\
& = & \prod_{a\in\mathcal{X}} 2^{n(P_{\boldsymbol{x}}(a)\log Q(a)-P_{\boldsymbol{x}}(a)\log P_{\boldsymbol{x}}(a)+P_{\boldsymbol{x}}(a)\log P_{\boldsymbol{x}}(a))} \\
& = & 2^{n\sum_{a\in\mathcal{X}}(-P_{\boldsymbol{x}}(a)\log\frac{P_{\boldsymbol{x}}(a)}{Q(a)}+P_{\boldsymbol{x}}(a)\log P_{\boldsymbol{x}}(a))} \\
& = & 2^{n(-D(P_{\boldsymbol{x}}\|Q)-H(P_{\boldsymbol{x}}))}. \\
& = & 2^{-n(H(P_{\boldsymbol{x}})+D(P_{\boldsymbol{x}}\|Q))}.
\end{array}$$

A Consequence

Corollary

If $\boldsymbol{x}$ is in the type class of Q, then

$$Q^n(\boldsymbol{x}) = 2^{-nH(Q)}.$$

- Suppose $\boldsymbol{x} \in T(Q)$.

  Then, we get

$$\begin{aligned} Q^n(\boldsymbol{x}) &= 2^{-n(H(P_{\boldsymbol{x}})+D(P_{\boldsymbol{x}}\|Q))} \\ &= 2^{-n(H(Q)+D(Q\|Q))} \\ &= 2^{-nH(Q)}. \end{aligned}$$

Example

- Consider rolling a fair die n (multiple of 6) times.

  It produces a particular sequence of length n, with probability

$$2^{-nH\left(\frac{1}{6},\frac{1}{6},\ldots,\frac{1}{6}\right)} = \left(\frac{1}{6}\right)^n.$$

- Suppose, next, that the die has a probability mass function

$$\left(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{12},\frac{1}{12},0\right).$$

  Then the probability of observing a particular sequence of length n (with n a multiple of 12) is precisely

$$2^{-nH\left(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{12},\frac{1}{12},0\right)}.$$

  This is a more interesting result.

The Size of a Type Class

- The exact size of $T(P)$ is a simple combinatorial problem.
- The size coincides with the number of ways of arranging $nP(a_1), nP(a_2), \ldots, nP(a_{|\mathcal{X}|})$ objects in a sequence.
- This number is

$$|T(P)| = \binom{n}{nP(a_1), nP(a_2), \ldots, nP(a_{|\mathcal{X}|})}.$$

- This value is hard to manipulate, so we derive simple exponential bounds on its value next.

Estimates of the Size of a Type Class

Theorem (Size of a Type Class $T(P)$)

For any type $P \in \mathcal{P}_n$,

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

- We first prove the upper bound.

  A type class must have probability ≤ 1.

  So we get

$$1 \geq P^n(T(P)) = \sum_{\boldsymbol{x} \in T(P)} P^n(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in T(P)} 2^{-nH(P)} = |T(P)| 2^{-nH(P)}.$$

  Thus, $|T(P)| \leq 2^{nH(P)}$.

Estimate of the Size of a Type Class (Lower Bound)

- We next work to establish the lower bound.

  We first prove that the type class $T(P)$ has the highest probability among all type classes under the probability distribution P,
$$P^n(T(P)) \geq P^n(T(\widehat{P})), \quad \text{for all } \widehat{P} \in \mathcal{P}_n.$$
  We lower bound the ratio of probabilities,
$$\begin{aligned} \frac{P^n(T(P))}{P^n(T(\widehat{P}))} &= \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\widehat{P})| \prod_{a\in\mathcal{X}} P(a)^{n\widehat{P}(a)}} \\ &= \frac{\binom{n}{nP(a_1),nP(a_2),\ldots,nP(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\widehat{P}(a_1),n\widehat{P}(a_2),\ldots,n\widehat{P}(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{n\widehat{P}(a)}} \\ &= \prod_{a\in\mathcal{X}} \frac{(n\widehat{P}(a))!}{(nP(a))!} P(a)^{n(P(a)-\widehat{P}(a))}. \end{aligned}$$

Estimate of the Size of a Type Class (Lower Bound Cont'd)

- We got
$$\frac{P^n(T(P))}{P^n(T(\widehat{P}))}=\prod_{a\in\mathcal{X}}\frac{(n\widehat{P}(a))!}{(nP(a))!}P(a)^{n(P(a)-\widehat{P}(a))}.$$
We use the simple bound $\frac{m!}{n!}\geq n^{m-n}$.
We obtain
$$\begin{array}{rcl}\frac{P^n(T(P))}{P^n(T(\widehat{P}))} & \geq & \prod_{a\in\mathcal{X}}(nP(a))^{n\widehat{P}(a)-nP(a)}P(a)^{n(P(a)-\widehat{P}(a))}\\ & = & \prod_{a\in\mathcal{X}}n^{n(\widehat{P}(a)-P(a))}\\ & = & n^{n(\sum_{a\in\mathcal{X}}\widehat{P}(a)-\sum_{a\in\mathcal{X}}P(a))}\\ & = & n^{n(1-1)}=1.\end{array}$$
Hence, $P^n(T(P))\geq P^n(T(\widehat{P}))$.

Estimate of the Size of a Type Class (Lower Bound Cont'd)

- The lower bound now follows easily from this result:

$$\begin{array}{rcl} 1 & = & \sum_{Q\in\mathcal{P}_n} P^n(T(Q)) \\ & \leq & \sum_{Q\in\mathcal{P}_n} \max_Q P^n(T(Q)) \\ & = & \sum_{Q\in\mathcal{P}_n} P^n(T(P)) \\ & \leq & (n+1)^{|\mathcal{X}|} P^n(T(P)) \\ & & \text{(by a previous theorem)} \\ & = & (n+1)^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in T(P)} P^n(\boldsymbol{x}) \\ & = & (n+1)^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in T(P)} 2^{-nH(P)} \\ & & \text{(by a previous theorem)} \\ & = & (n+1)^{|\mathcal{X}|}|T(P)|2^{-nH(P)}. \end{array}$$

Binary Alphabet (Upper Bound)

- In the binary case, the type is defined by the number k of 1's in the sequence, and the size of the type class is $\binom{n}{k}$.

  Claim: $\frac{1}{n+1}2^{nH\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH\left(\frac{k}{n}\right)}$.

  We first prove the upper bound.

  By the binomial formula, for any $0 \leq p \leq 1$,

  $$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

  All the terms of the sum are positive and sum up to 1.

  Thus, each of the terms is less than 1.

Binary Alphabet (Upper Bound Cont'd)

- Set $p = \frac{k}{n}$ and consider the k-th term

$$\begin{array}{rcl} 1 & \geq & \binom{n}{k}(\frac{k}{n})^k(1-\frac{k}{n})^{n-k} \\ & = & \binom{n}{k}2^{k\log\frac{k}{n}+(n-k)\log\frac{n-k}{n}} \\ & = & \binom{n}{k}2^{n(\frac{k}{n}\log\frac{k}{n}+\frac{n-k}{n}\log\frac{n-k}{n})} \\ & = & \binom{n}{k}2^{-nH(\frac{k}{n})}. \end{array}$$

Hence

$$\binom{n}{k} \leq 2^{nH(\frac{k}{n})}.$$

Binary Alphabet (Lower Bound)

- For the lower bound, let S be a random variable with a binomial distribution with parameters n and p.
  The most likely value of S is $S = \langle np \rangle$.
  To see this, compute
  $$\frac{P(S=i+1)}{P(S=i)} = \frac{\binom{n}{i+1} p^{i+1}(1-p)^{n-i-1}}{\binom{n}{i} p^i (1-p)^{n-i}} = \frac{n-i}{i+1}\frac{p}{1-p}.$$
  Suppose, first, that $i < np$.
  Then $i - ip < np - ip + p$. Hence, $i(1-p) < (n-(i-1))p$.
  This gives $\frac{P(S=i)}{P(S=i-1)} = \frac{(n-(i-1))p}{i(1-p)} > 1$.
  Therefore, $P(S=i-1) < P(S=i)$.
  Suppose, next, that $np < i$.
  Then $np - ip < i + 1 - ip - p$. Hence, $(n-i)p < (i+1)(1-p)$.
  This gives $\frac{P(S=i+1)}{P(S=i)} = \frac{(n-i)p}{(i+1)(1-p)} < 1$.
  Therefore, $P(S=i+1) < P(S=i)$.

Binary Alphabet (Lower Bound Cont'd)

- Since there are $n+1$ terms in the binomial sum,

$$\begin{array}{rcl} 1 & = & \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \\ & \leq & (n+1) \max_k \binom{n}{k} p^k (1-p)^{n-k} \\ & = & (n+1) \binom{n}{\langle np \rangle} p^{\langle np \rangle} (1-p)^{n-\langle np \rangle}. \end{array}$$

Now let $p = \frac{k}{n}$. Then we have $1 \leq (n+1)\binom{n}{k}(\frac{k}{n})^k(1-\frac{k}{n})^{n-k}$.

By the previous arguments this is equivalent to $\frac{1}{n+1} \leq \binom{n}{k} 2^{-nH(\frac{k}{n})}$.

Rewriting,

$$\binom{n}{k} \geq \frac{1}{n+1} 2^{nH(\frac{k}{n})}.$$

Combining the two results, $\binom{n}{k} \doteq 2^{nH(\frac{k}{n})}$.

Probability of Type Class

Theorem (Probability of Type Class)

For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P\|Q)}$ to first order in the exponent. More precisely,

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

- We have

$$\begin{array}{rcl} Q^n(T(P)) & = & \sum_{\boldsymbol{x} \in T(P)} Q^n(\boldsymbol{x}) \\ & = & \sum_{\boldsymbol{x} \in T(P)} 2^{-n(D(P\|Q)+H(P))} \\ & & \text{(by a previous theorem)} \\ & = & |T(P)| 2^{-n(D(P\|Q)+H(P))}. \end{array}$$

  Now, we ue the bounds on $|T(P)|$ derived previously.
  We get $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)}$.

Summary of Results

- We can summarize the basic theorems concerning types in four equations.

$$\begin{aligned} |\mathcal{P}_n| &\leq (n+1)^{|\mathcal{X}|}; \\ Q^n(\boldsymbol{x}) &= 2^{-n(D(P_{\boldsymbol{x}}\|Q)+H(P_{\boldsymbol{x}}))}; \\ |T(P)| &\doteq 2^{nH(P)}; \\ Q^n(T(P)) &\doteq 2^{-nD(P\|Q)}. \end{aligned}$$

- These equations state that:
  - There are only a polynomial number of types;
  - There are an exponential number of sequences of each type.
- We also have:
  - An exact formula for the probability of any sequence of type P under distribution Q;
  - An approximate formula for the probability of a type class.

Subsection 2

Law of Large Numbers

Typical Sequences

- Given an $\epsilon > 0$, we define a **typical set** T_Q^ϵ of sequences for the distribution Q^n by

$$T_Q^\epsilon = \{x^n : D(P_{x^n} \| Q) \leq \epsilon\}.$$

Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim P(x)$. Then

$$\Pr\{D(P_{x^n} \| P) > \epsilon\} \leq 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)}.$$

Consequently, $D(P_{x^n} \| P) \to 0$ with probability 1.

Typical Sequences (Cont'd)

- The probability that x^n is not typical is

$$\begin{array}{rcl} 1 - Q^n(T_Q^\epsilon) & = & \sum_{P:D(P\|Q)>\epsilon} Q^n(T(P)) \\ & \leq & \sum_{P:D(P\|Q)>\epsilon} 2^{-nD(P\|Q)} \\ & \leq & \sum_{P:D(P\|Q)>\epsilon} 2^{-n\epsilon} \\ & \leq & (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \\ & = & 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)}. \end{array}$$

Typical Sequences (Cont'd)

- Summing over n, we get

$$\sum_{n=1}^{\infty} \Pr\{D(P_{x^n}\|P) > \epsilon\} < \infty.$$

Thus, the expected number of occurrences of the event

$$\{D(P_{x^n}\|P) > \epsilon\}$$

for all n is finite.

By the Borel-Cantelli Lemma, this implies that the actual number of such occurrences is also finite with probability 1.

Hence, $D(P_{x^n}\|P) \to 0$ with probability 1.

Strongly Typical Sequences

Definition

We define the **strongly typical set** $A_\epsilon^{*(n)}$ to be the set of sequences in $\mathcal{X}^n$ for which the sample frequencies are close to the true values:

$$A^{*(n)} = \left\{ \boldsymbol{x} \in \mathcal{X}^n : \begin{array}{ll} \left|\frac{1}{n}N(a|\boldsymbol{x}) - P(a)\right| < \frac{\epsilon}{|\mathcal{X}|}, & \text{if } P(a) > 0 \\ N(a|\boldsymbol{x}) = 0, & \text{if } P(a) = 0 \end{array} \right\}.$$

- Hence, the typical set consists of sequences whose type does not differ from the true probabilities by more than $\frac{\epsilon}{|\mathcal{X}|}$ in any component.
- By the Strong Law of Large Numbers, the probability of the strongly typical set goes to 1 as $n \to \infty$.

Subsection 3

Universal Source Coding

Unknown Source Distributions

- Huffman coding compresses an i.i.d. source with a known distribution $p(x)$ to its entropy limit $H(X)$.
- If the code is designed for some incorrect distribution $q(x)$, a penalty of $D(p\|q)$ is incurred.
- Thus, Huffman coding is sensitive to the assumed distribution.
- Suppose, now, the true distribution $p(x)$ is unknown.
- We describe a universal code of rate R that suffices to describe every i.i.d. source with entropy $H(X) < R$.
- The idea is based on the *method of types*.

The Idea Behind the Universal Coding

- There are $2^{nH(P)}$ sequences of type P.
- There are only a polynomial number of types with denominator n.
- So an enumeration of all sequences x^n with type P_{x^n}, such that $H(P_{x^n}) < R$, will require roughly nR bits.
- By describing all such sequences, we are prepared to describe any sequence that is likely to arise from any distribution Q having entropy $H(Q) < R$.

Universal Codes

Definition

A **fixed-rate block code of rate** R for a source $X_1, X_2, \ldots, X_n$, which has an unknown distribution Q, consists of two mappings.

- The **encoder** $f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$;
- The **decoder** $\phi_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n$.

The probability of error for the code with respect to the distribution Q is

$$P_e^{(n)} = Q^n(X^n : \phi_n(f_n(X^n)) \neq X^n).$$

Definition

A rate R block code for a source will be called **universal** if the functions f_n and ϕ_n do not depend on the distribution Q and if $P_e^{(n)} \to 0$ as $n \to \infty$, if $R > H(Q)$.

Universal Encoding Scheme

Theorem

There exists a sequence of $(2^{nR}, n)$ universal source codes, such that $P_e^{(n)} \to 0$, for every source Q such that $H(Q) < R$.

- Fix the rate R for the code. Let $R_n = R - |\mathcal{X}|\frac{\log(n+1)}{n}$. Consider the set of sequences $A = \{x \in \mathcal{X}^n : H(P_x) \leq R_n\}$. Then

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

Universal Encoding Scheme (Encoding and Decoding)

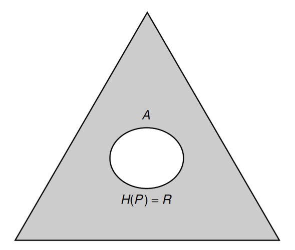
- By indexing the elements of A, we define the encoding function f_n as

$$f_n(\boldsymbol{x}) = \begin{cases} \text{index of } \boldsymbol{x} \text{ in } A, & \text{if } \boldsymbol{x} \in A \\ 0, & \text{otherwise} \end{cases}.$$

  The decoding function maps each index onto the corresponding element of A.

  - All the elements of A are recovered correctly;
  - All the remaining sequences result in an error.

  The set of sequences that are recovered correctly is illustrated in the figure.



Universality

- We now show that this encoding scheme is universal.

  Assume that the distribution of $X_1, X_2, \ldots, X_n$ is Q and $H(Q) < R$.

  Then the probability of decoding error is given by

$$\begin{array}{rcl} P_e^{(n)} & = & 1 - Q^n(A) \\ & = & \sum_{P : H(P) > R_n} Q^n(T(P)) \\ & \leq & (n+1)^{|\mathcal{X}|} \max_{P : H(P) > R_n} Q^n(T(P)) \\ & \leq & (n+1)^{|\mathcal{X}|} 2^{-n \min_{P : H(P) > R_n} D(P \| Q)}. \end{array}$$

  Now, $R_n \uparrow R$ and $H(Q) < R$.

  So, there exists n_0, such that, for all $n \geq n_0$, $R_n > H(Q)$.

  Then, for $n \geq n_0$, $\min_{P : H(P) > R_n} D(P \| Q)$ must be greater than 0.

  So the probability of error $P_e^{(n)} \overset{n \to \infty}{\longrightarrow} 0$ exponentially fast.

Universal Encoding Scheme (Conclusion)

- Suppose the distribution Q is such that $H(Q) > R$.

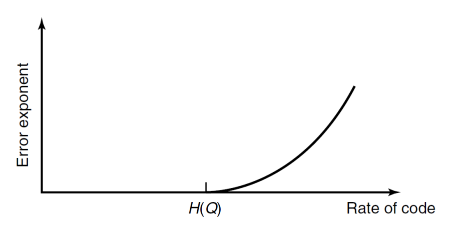
  Then with high probability the sequence will have a type outside A.

  Hence, in such cases the probability of error is close to 1.

  The exponent in the probability of error is

$$D^*_{R,Q} = \min_{P:H(P)>R} D(P\|Q).$$

  It is illustrated in the figure.



Huffman Codes versus Universal Codes

- Why it is ever necessary to use Huffman codes, which are specific to a probability distribution?
- What do we lose in using a universal code?
- Universal codes need a longer block length to obtain the same performance as a code designed specifically for the probability distribution.
- The penalty for this increase in block length is the increased complexity of the encoder and decoder.
- Hence, a distribution specific code is best if one knows the distribution of the source.

Subsection 4

Large Deviation Theory

Subject of Large Deviation Theory

- Suppose $X_1, X_2, \ldots, X_n$ are drawn i.i.d. $\sim$ Bernoulli $\left(\frac{1}{3}\right)$.
- We want to find the probability that $\frac{1}{n}\sum X_i$ is near $\frac{1}{3}$.
- This is a small deviation (from the expected outcome) and the probability is near 1.
- How about the probability that $\frac{1}{n}\sum X_i$ is greater than $\frac{3}{4}$?
- This is a large deviation, and the probability is exponentially small.
- One estimate of the exponent uses the central limit theorem.
- But this is a poor approximation for more than a few standard deviations.
- Note that $\frac{1}{n}\sum X_i = \frac{3}{4}$ is equivalent to $P_{\boldsymbol{X}} = (\frac{1}{4}, \frac{3}{4})$.
- Thus, the probability that $\overline{X}_n$ is near $\frac{3}{4}$ is the probability that type P_X is near $(\frac{3}{4}, \frac{1}{4})$.
- The probability of this large deviation is $\approx 2^{-nD\left(\left(\frac{3}{4},\frac{1}{4}\right)\|\left(\frac{1}{3},\frac{2}{3}\right)\right)}$.

The Setup and the Goal

- Let E be a subset of the set of probability mass functions.

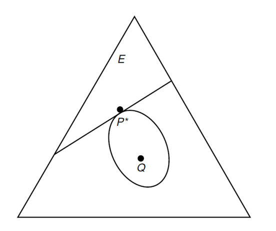
  Example: E may be the set of probability mass functions with mean μ.
- With a slight abuse of notation, we write

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) = \sum_{\boldsymbol{x} : P_{\boldsymbol{x}} \in E \cap \mathcal{P}_n} Q^n(\boldsymbol{x}).$$

- If E contains a relative entropy neighborhood of Q, then by the weak law of large numbers, $Q^n(E) \to 1$.
- If E does not contain Q or a neighborhood of Q, then by the weak law of large numbers, $Q^n(E) \to 0$ exponentially fast.
- We use the method of types to calculate the exponent.

Example

- Assume that, by observation, we find that the sample average of $g(X)$ is greater than or equal to α, i.e., $\frac{1}{n}\sum_i g(x_i) \geq \alpha$. This event is equivalent to the event $P_X \in E \cap \mathcal{P}_n$, where
$$E = \left\{P : \sum_{a \in \mathcal{X}} g(a)P(a) \geq \alpha\right\}.$$
Indeed, we have
$$\begin{array}{rcl} \frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha & \Leftrightarrow & \sum_{a \in \mathcal{X}} P_X(a)g(a) \geq \alpha \\ & \Leftrightarrow & P_X \in E \cap \mathcal{P}_n. \end{array}$$
Thus,
$$\begin{array}{rcl} \Pr(\frac{1}{n}\sum_{i=1}^n g(X_i) \geq \alpha) & = & Q^n(E \cap \mathcal{P}_n) \\ & = & Q^n(E). \end{array}$$
Here E is a half space in the space of probability vectors, as shown in the figure.



Sanov's Theorem

Theorem (Sanov's Theorem)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^* \| Q)},$$

where $P^* = \operatorname{argmin}_{P \in E} D(P \| Q)$ is the distribution in E that is closest to Q in relative entropy.

If, in addition, the set E is the closure of its interior, then

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

Proof of Sanov's Theorem (Upper Bound)

- We first prove the upper bound:

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

where the last inequality follows from a previous theorem.

Note that P^* need not be a member of $\mathcal{P}_n$.

Proof of Sanov's Theorem (Asymptotic Bahavior)

- We now look for a "nice" set E, so that, for all large n, there is a distribution in $E \cap \mathcal{P}_n$ close to P^*.

  Assume E is the closure of its interior (so the interior is nonempty).

  $\bigcup_n \mathcal{P}_n$ is dense in the set of all distributions.

  Hence, $E \cap \mathcal{P}_n$ is nonempty, for all $n \geq n_0$, for some n_0.

  Find a sequence $P_n \in E \cap \mathcal{P}_n$, such that $D(P_n \| Q) \to D(P^* \| Q)$.

  Then, for each $n \geq n_0$,

$$Q^n(E) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \geq Q^n(T(P_n)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n \| Q)}.$$

  Consequently,

$$\begin{array}{rcl} \liminf \frac{1}{n} \log Q^n(E) & \geq & \liminf \left(-\frac{|\mathcal{X}| \log (n+1)}{n} - D(P_n \| Q)\right) \\ & = & -D(P^* \| Q). \end{array}$$

  Combining this with the upper bound establishes the theorem.

Subsection 5

Examples of Sanov's Theorem

Minimizing the Relative Entropy

- Suppose that we wish to find

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) \geq \alpha_j, \quad j = 1, 2, \ldots, k\right\}.$$

- The set E is defined as

$$E = \left\{P : \sum_{a} P(a) g_j(a) \geq \alpha_j, j = 1, 2, \ldots, k\right\}.$$

- To find the closest distribution in E to Q, we minimize $D(P\|Q)$ subject to the constraints above.
- Using Lagrange multipliers, we construct the functional

$$J(P) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} + \sum_{i} \lambda_i \sum_{x} P(x) g_i(x) + \nu \sum_{x} P(x).$$

Minimizing the Relative Entropy

- We obtained the functional

$$J(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_i \lambda_i \sum_x P(x) g_i(x) + \nu \sum_x P(x).$$

- We then differentiate and calculate the closest distribution to Q to be of the form

$$P^*(x) = \frac{Q(x) e^{\sum_i \lambda_i g_i(x)}}{\sum_{a \in \mathcal{X}} Q(a) e^{\sum_i \lambda_i g_i(a)}},$$

where the constants λ_i are chosen to satisfy the constraints.

- Note that if Q is uniform, P^* is the maximum entropy distribution.

Example: Dice

- Suppose that we toss a fair die n times.

  We compute the probability that the average of the throws is greater than or equal to 4.

  Let P^* be the distribution that minimizes $D(P\|Q)$ over all distributions P that satisfy $\sum_{i=1}^6 iP(i) \geq 4$.

  By Sanov's theorem, it follows that

$$Q^n(E) \doteq 2^{-nD(P^*\|Q)}.$$

  From the minimization formula it follows that P^* has the form

$$P^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^6 2^{\lambda i}},$$

  with λ chosen so that $\sum iP^*(i) = 4$.

Example: Dice (Cont'd)

- We got
$$P^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda i}},$$
with λ chosen so that $\sum iP^*(i) = 4$.

  Solving numerically, we obtain:
  - $\lambda = 0.2519$;
  - $P^* = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468)$.

  Therefore,
$$D(P^* \| Q) = 0.0624 \text{ bit}.$$

  Thus, the probability that the average of 10000 throws is greater than or equal to 4 is $\approx 2^{-624}$.

Example: Coins

- Suppose that we have a fair coin and want to estimate the probability of observing more than 700 heads in a series of 1000 tosses.

  The problem is like in the preceding example.

  The probability is

  $$P(\overline{X}_n \geq 0.7) \doteq 2^{-nD(P\|Q)},$$

  where:

  - P^* is the $(0.7, 0.3)$ distribution;
  - Q is the $(0.5, 0.5)$ distribution.

  So we get

  $$D(P\|Q) = 1 - H(P^*) = 1 - H(0.7) = 0.119.$$

  Thus, the probability of 700 or more heads in 1000 trials is approximately 2^{-119}.

Example: Mutual Dependence

- Let $Q(x, y)$ be a given joint distribution.

  Let $Q_0(x, y) = Q(x)Q(y)$ be the associated product distribution formed from the marginals of Q.

  We wish to know the likelihood that a sample drawn according to Q_0 will "appear" to be jointly distributed according to Q.

  Let (X_i, Y_i) be i.i.d. $\sim Q_0(x, y) = Q(x)Q(y)$.

  We define joint typicality as we did before.

  I.e., (x^n, y^n) is **jointly typical with respect to a joint distribution** $Q(x, y)$ iff the sample entropies are close to their true values:

$$\left|-\frac{1}{n}\log Q(x^n) - H(X)\right| \leq \epsilon,$$
$$\left|-\frac{1}{n}\log Q(y^n) - H(Y)\right| \leq \epsilon,$$
$$\left|-\frac{1}{n}\log Q(x^n, y^n) - H(X, Y)\right| \leq \epsilon.$$

Example: Mutual Dependence (Cont'd)

- We wish to calculate the probability (under the product distribution) of seeing a pair (x^n, y^n) that looks jointly typical of Q.

  Thus, (x^n, y^n) are jointly typical with respect to $Q(x, y)$ if $P_{x^n,y^n} \in E \cap \mathcal{P}_n(X, Y)$, where

$$\begin{aligned} E = \{P(x, y) : \quad & |-\textstyle\sum_{x,y} P(x, y) \log Q(x) - H(X)| \leq \epsilon, \\ & |-\textstyle\sum_{x,y} P(x, y) \log Q(y) - H(Y)| \leq \epsilon, \\ & |-\textstyle\sum_{x,y} P(x, y) \log Q(x, y) - H(X, Y)| \leq \epsilon\}. \end{aligned}$$

  Using Sanov's Theorem, the probability is $Q_0^n(E) \doteq 2^{-nD(P\|Q_0)}$, where P^* is the distribution satisfying the constraints that is closest to Q_0 in relative entropy.

  In this case, as $\epsilon \to 0$, it can be verified that P^* is the joint distribution Q, and Q_0 is the product distribution.

  So the probability is $2^{-nD(Q(x,y)\|Q(x)Q(y))} = 2^{-nI(X;Y)}$.

  This coincides with the result derived previously for the joint AEP.

Subsection 6

Conditional Limit Theorem

Review

- It has been shown that the probability of a set of types under a distribution Q is determined essentially by the probability of the closest element of the set to Q.
- This probability is
$$2^{-nD^*}$$
to first order in the exponent, where $D^* = \min_{P \in E} D(P\|Q)$.
- This follows from the following considerations.
  - The probability of the set of types is the sum of the probabilities of each type, which is bounded by the largest term times the number of terms.
  - The number of terms is polynomial in the length of the sequences.
  - So the sum is equal to the largest term to first order in the exponent.

A New Goal

- We strengthen the argument to show that, not only is the probability of the set E essentially the same as the probability of the closest type P^*, but also that the total probability of other types that are far away from P^* is negligible.
- This implies that with very high probability, the type observed is close to P^*.
- This fact is referred to as a conditional limit theorem.

A “Pythagorean” Theorem

- Since $D(P\|Q)$ is not a metric, many of the intuitive properties of distance are not valid for $D(P\|Q)$.
- The next theorem shows a sense in which $D(P\|Q)$ behaves like the square of the Euclidean metric.

Theorem

For a closed convex set $E \subseteq \mathcal{P}$ and distribution $Q \notin E$, let $P^* \in E$ be the distribution that achieves the minimum distance to Q, i.e., $D(P^*\|Q) = \min_{P\in E} D(P\|Q)$. Then, for all $P \in E$,

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q).$$

Note: The main use of this theorem is as follows.

Suppose a sequence $P_n \in E$ satisfies $D(P_n\|Q) \to D(P^*\|Q)$.

Then, by the Pythagorean Theorem, $D(P_n\|P^*) \to 0$ as well.

Proof of the Theorem

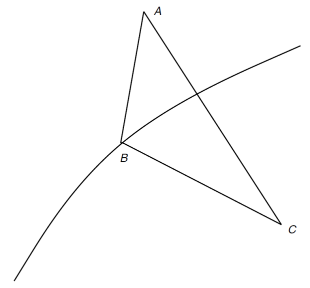
- Consider any $P \in E$. Let
$$P_\lambda = \lambda P + (1-\lambda)P^*.$$
Then $P_\lambda \to P^*$ as $\lambda \to 0$.
Also, since E is convex, $P_\lambda \in E$, for $0 \le \lambda \le 1$.
$D(P^* \| Q)$ is the minimum of $D(P_\lambda \| Q)$ along the path $P^* \to P$.
So the derivative of $D(P_\lambda \| Q)$ as a function of λ is nonnegative at $\lambda = 0$.
To take advantage of this observation, compute:
$$D_\lambda = D(P_\lambda \| Q) = \sum P_\lambda(x) \log \frac{P_\lambda(x)}{Q(x)};$$
$$\frac{dD_\lambda}{d\lambda} = \sum \left((P(x) - P^*(x)) \log \frac{P_\lambda(x)}{Q(x)} + (P(x) - P^*(x)) \right).$$

Proof of the Theorem (Cont'd)

- We have
$$\frac{dD_\lambda}{d\lambda}=\sum\left((P(x)-P^*(x))\log\frac{P_\lambda(x)}{Q(x)}+(P(x)-P^*(x))\right).$$
Set $\lambda=0$. Then $P_\lambda=P^*$. Moreover, $\sum P(x)=\sum P^*(x)=1$. Now we get
$$\begin{array}{rcl}0&\leq&(\frac{dD_\lambda}{d\lambda})_{\lambda=0}\\&=&\sum(P(x)-P^*(x))\log\frac{P^*(x)}{Q(x)}\\&=&\sum P(x)\log\frac{P^*(x)}{Q(x)}-\sum P^*(x)\log\frac{P^*(x)}{Q(x)}\\&=&\sum P(x)\log\frac{P(x)}{Q(x)}\frac{P^*(x)}{P(x)}-\sum P^*(x)\log\frac{P^*(x)}{Q(x)}\\&=&D(P\|Q)-D(P\|P^*)-D(P^*\|Q).\end{array}$$

Illustration

- The relative entropy $D(P\|Q)$ behaves like the square of the Euclidean distance.



- Suppose we have a convex set E in $\mathbb{R}^n$.
- Let:
  - A be a point outside the set;
  - B be the point in the set closest to A;
  - C be any other point in the set.
- The angle between the lines BA and BC must be obtuse.
- This implies that

$$\ell_{AC}^2 \geq \ell_{AB}^2 + \ell_{BC}^2.$$

The $\mathcal{L}_1$ Norm

Definition

The $\mathcal{L}_1$ **distance** between any two distributions is defined as

$$\|P_1 - P_2\|_1 = \sum_{a \in \mathcal{X}} |P_1(a) - P_2(a)|.$$

- Let A be the set on which $P_1(x) > P_2(x)$. Then

$$\begin{array}{rcl} \|P_1 - P_2\|_1 & = & \sum_{x \in \mathcal{X}} |P_1(x) - P_2(x)| \\ & = & \sum_{x \in A}(P_1(x) - P_2(x)) + \sum_{x \in A^c}(P_2(x) - P_1(x)) \\ & = & P_1(A) - P_2(A) + P_2(A^c) - P_1(A^c) \\ & = & P_1(A) - P_2(A) + 1 - P_2(A) - 1 + P_1(A) \\ & = & 2(P_1(A) - P_2(A)). \end{array}$$

The Variational Distance

- Let, again, P_1, P_2 be distributions.
- The **variational distance** between P_1 and P_2 is defined by

$$\max_{B \subseteq \mathcal{X}} (P_1(B) - P_2(B)).$$

- We have

$$\begin{aligned} \max_{B \subseteq \mathcal{X}} (P_1(B) - P_2(B)) &= P_1(A) - P_2(A) \\ &= \frac{\|P_1 - P_2\|_1}{2}. \end{aligned}$$

Convergence in Relative Entropy and in $\mathcal{L}_1$ Norm

Lemma

We have

$$D(P_1 \| P_2) \geq \frac{1}{2 \ln 2} \|P_1 - P_2\|_1^2.$$

- We first prove it for the binary case.

  Consider two binary distributions with parameters p and q, $p \geq q$.

  We show

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \frac{4}{2 \ln 2}(p-q)^2.$$

  The difference $g(p,q)$ between the two sides is

$$g(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - \frac{4}{2 \ln 2}(p-q)^2.$$

Convergence in Relative Entropy and in $\mathcal{L}_1$ Norm (Cont'd)

- We defined $g(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - \frac{4}{2\ln 2}(p-q)^2$. Then

$$\begin{array}{rcl}\dfrac{dg(p,q)}{dq} & = & -\dfrac{p}{q\ln 2} + \dfrac{1-p}{(1-q)\ln 2} - \dfrac{4}{2\ln 2}2(q-p) \\ & = & \dfrac{q-p}{q(1-q)\ln 2} - \dfrac{4}{\ln 2}(q-p) \\ & \leq & 0. \quad \left(q(1-q) \leq \dfrac{1}{4}, q \leq p\right)\end{array}$$

For $q = p$, $g(p,q) = 0$.

Hence, for $q \leq p$, $g(p,q) \geq 0$.

This proves the lemma for the binary case.

Convergence in Relative Entropy and in $\mathcal{L}_1$ Norm (Cont'd)

- For the general case, for any two distributions P_1 and P_2, let
$$A = \{x : P_1(x) > P_2(x)\}.$$
Define a binary random variable $Y = \phi(X)$, the indicator of A.
Let $\widehat{P}_1$ and $\widehat{P}_2$ be the distributions of Y.
Thus, $\widehat{P}_1$, $\widehat{P}_2$ correspond to the quantized versions of P_1, P_2.
By the Data-Processing Inequality applied to relative entropies (which is proved in the same way as the Data-Processing Inequality for mutual information), we have
$$\begin{array}{rcl} D(P_1\|P_2) & \geq & D(\widehat{P}_1\|\widehat{P}_2) \\ & \geq & \frac{4}{2\ln 2}(P_1(A) - P_2(A))^2 \\ & = & \frac{1}{2\ln 2}\|P_1 - P_2\|_1^2. \end{array}$$

Outline of Proof of Conditional Limit Theorem

- The essential idea is that the probability of a type under Q depends exponentially on the distance of the type from Q.
- Hence types that are farther away are exponentially less likely to occur.
- We divide the set of types in E into two categories.
  - Those at about the same distance from Q as P^*;
  - Those a distance 2δ farther away.
- The second set has exponentially less probability than the first.
- Hence, the first set has a conditional probability tending to 1.
- We then use the Pythagorean theorem to establish that all the elements in the first set are close to P^*, which will establish the theorem.

Conditional Limit Theorem

Theorem (Conditional Limit Theorem)

Let E be a closed convex subset of $\mathcal{P}$ and let Q be a distribution not in E. Let $X_1, X_2, \ldots, X_n$ be discrete random variables drawn i.i.d. $\sim Q$. Let P^* achieve $\min_{P \in E} D(P \| Q)$. Then

$$\Pr(X_1 = a | P_{X^n} \in E) \to P^*(a) \text{ in probability as } n \to \infty,$$

i.e., the conditional distribution of X_1, given that the type of the sequence is in E, is close to P^* for large n.

Example of Conditional Limit Theorem

- Suppose X_i i.i.d. $\sim Q$.

  Let $P^*(a)$ minimize $D(P\|Q)$, over P satisfying $\sum P(a)a^2 \geq \alpha$.

  Then
  $$\Pr\left\{X_1 = a : \frac{1}{n}\sum X_i^2 \geq \alpha\right\} \to P^*(a).$$

  This minimization results in
  $$P^*(a) = Q(a)\frac{e^{\lambda a^2}}{\sum aQ(a)e^{\lambda a^2}},$$

  where λ is chosen to satisfy $\sum P^*(a)a^2 = \alpha$.

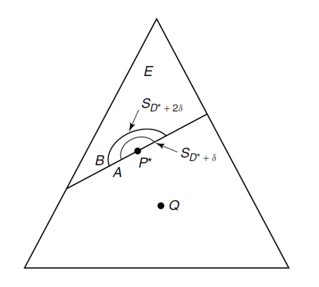
  Thus, the conditional distribution on X_1 given a constraint on the sum of the squares is a (normalized) product of the original probability mass function and the maximum entropy probability mass function (which in this case is Gaussian).

Proof of the Conditional Limit Theorem

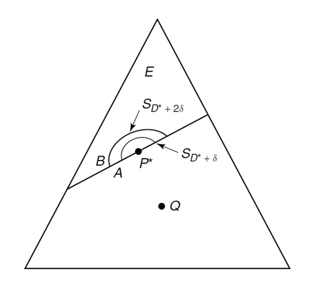
- Define the sets
$$S_t = \{P \in \mathcal{P} : D(P\|Q) \leq t\}.$$
The set S_t is convex, since $D(P\|Q)$ is a convex function of P.
Let
$$D^* = D(P^*\|Q) = \min_{P \in E} D(P\|Q).$$
Then P^* is unique, since $D(P\|Q)$ is strictly convex in P.

Now define the sets:
  - $A = S_{D^*+2\delta} \cap E$;
  - $B = E - (S_{D^*+2\delta} \cap E)$.

Thus, $A \cup B = E$.



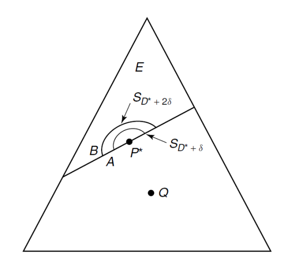
Proof of the Conditional Limit Theorem (Cont'd)



Then we have

$$\begin{aligned} Q^n(B) &= \sum_{P\in E\cap\mathcal{P}_n:D(P\|Q)>D^*+2\delta} Q^n(T(P)) \\ &\leq \sum_{P\in E\cap\mathcal{P}_n:D(P\|Q)>D^*+2\delta} 2^{-nD(P\|Q)} \\ &\leq \sum_{P\in E\cap\mathcal{P}_n:D(P\|Q)>D^*+2\delta} 2^{-n(D^*+2\delta)} \\ &\leq (n+1)^{|\mathcal{X}|}2^{-n(D^*+2\delta)}. \end{aligned}$$

Proof of the Conditional Limit Theorem (Cont'd)



We also have

$$\begin{array}{rcl} Q^n(A) & \geq & Q^n(S_{D^*+\delta} \cap E) \\ & = & \sum_{P \in E \cap \mathcal{P}_n : D(P\|Q) \leq D^*+\delta} Q^n(T(P)) \\ & \geq & \sum_{P \in E \cap \mathcal{P}_n : D(P\|Q) \leq D^*+\delta} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \\ & \geq & \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}, \text{ for } n \text{ sufficiently large,} \end{array}$$

since the sum is greater than one of the terms, and for sufficiently large n, there exists at least one type in $S_{D^*+\delta} \cap E \cap \mathcal{P}_n$.

Proof of the Conditional Limit Theorem (Cont'd)

- Then, for n sufficiently large,

$$\begin{array}{rcl}\Pr(P_{X^n} \in B | P_{X^n} \in E) & = & \dfrac{Q^n(B \cap E)}{Q^n(E)} \\ & \leq & \dfrac{Q^n(B)}{Q^n(A)} \\ & \leq & \dfrac{(n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}} \\ & = & (n+1)^{2|\mathcal{X}|} 2^{-n\delta} \\ & \overset{n \to \infty}{\to} & 0. \end{array}$$

Hence, the conditional probability of B goes to 0 as $n \to \infty$.

So the conditional probability of A goes to 1.

Proof of the Conditional Limit Theorem (Cont'd)

- We show that all members of A are close to P^* in relative entropy.

  For all members of A, $D(P\|Q) \leq D^* + 2\delta$.

  Hence by the "Pythagorean" Theorem,

  $$D(P\|P^*) + D(P^*\|Q) \leq D(P\|Q) \leq D^* + 2\delta.$$

  Since $D(P^*\|Q) = D^*$,

  $$D(P\|P^*) \leq 2\delta.$$

  We know that $P_{\boldsymbol{X}} \in A$ implies that $D(P_{\boldsymbol{X}}\|Q) \leq D^* + 2\delta$.

  By what was just shown, $D(P_{\boldsymbol{X}}\|P^*) \leq 2\delta$.

Proof of the Conditional Limit Theorem (Conclusion)

- Since $\Pr\{P_{X^n} \in A | P_{X^n} \in E\} \to 1$, it follows that, as $n \to \infty$,

  $$\Pr(D(P_{X^n} \| P^*) \leq 2\delta | P_{X^n} \in E) \to 1.$$

  By a previous lemma, the fact that the relative entropy is small implies that the $\mathcal{L}_1$ distance is small.

  This, in turn, implies that

  $$\max_{a \in \mathcal{X}} |P_{X^n}(a) - P^*(a)|$$

  is small.

  Thus, as $n \to \infty$,

  $$\Pr(|P_{X^n}(a) - P^*(a)| \geq \epsilon | P_{X^n} \in E) \to 0.$$

  Alternatively, this can be written as $\Pr(X_1 = a | P_{X^n} \in E) \to P^*(a)$ in probability, $a \in \mathcal{X}$.

A Strengthening of the Theorem

- In this theorem we have only proved that the marginal distribution goes to P^* as $n \to \infty$.
- Using a similar argument, we can prove the stronger statement

$$\Pr(X_1 = a_1, X_2 = a_2, \ldots, X_m = a_m | P_{X^n} \in E) \to \prod_{i=1}^{m} P^*(a_i)$$

  in probability.
- This holds for fixed m as $n \to \infty$.
- The result is not true for $m = n$, since there are end effects.
- Given that the type of the sequence is in E, the last elements of the sequence can be determined from the remaining elements, and the elements are no longer independent.
- The Conditional Limit Theorem states that the first few elements are asymptotically independent with common distribution P^*.

Example

- Consider the case when n fair dice are rolled.
- Suppose that the sum of the outcomes exceeds $4n$.
- By the Conditional Limit Theorem, the probability that the first die shows a number $a \in \{1, 2, \ldots, 6\}$ is approximately $P^*(a)$, where $P^*(a)$ is the distribution in E that is closest to the uniform distribution, where $E = \{P : \sum P(a)a \geq 4\}$.
- This is the maximum entropy distribution given by
$$P^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda i}},$$
with λ chosen so that $\sum iP^*(i) = 4$.
- Here P^* is the conditional distribution on the first (or any other) die.
- Apparently, the first few dice inspected will behave as if they are drawn independently according to an exponential distribution.

Subsection 7

Hypothesis Testing

The Hypothesis Testing Problem

- The **hypothesis testing problem** in statistics is to decide between two alternative explanations for the data observed.
- In the simplest case, we have to decide between two i.i.d. distributions.

Problem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$. We consider two hypotheses:

H_1: $Q = P_1$.

H_2: $Q = P_2$.

The Setup

- Consider the general decision function $g(x_1, x_2, \ldots, x_n)$, where:
  - $g(x_1, x_2, \ldots, x_n) = 1$ means that H_1 is accepted;
  - $g(x_1, x_2, \ldots, x_n) = 2$ means that H_2 is accepted.
- Since the function takes on only two values, the test can also be specified by specifying the set A over which $g(x_1, x_2, \ldots, x_n)$ is 1.
- The complement of A is the set where $g(x_1, x_2, \ldots, x_n) = 2$.
- We define the two probabilities of error.

$$\begin{array}{rcl} \alpha & = & \Pr(g(X_1, X_2, \ldots, X_n) = 2 | H_1 \text{ true}) = P_1^n(A^c); \\ \beta & = & \Pr(g(X_1, X_2, \ldots, X_n) = 1 | H_2 \text{ true}) = P_2^n(A). \end{array}$$

- We wish to minimize both probabilities, but there is a tradeoff.
- Thus, we minimize one of the probabilities of error subject to a constraint on the other probability of error.

Neyman-Pearson Lemma for Optimum Test

Theorem (Neyman-Pearson Lemma)

Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according to probability mass function Q. Consider the decision problem corresponding to hypotheses $Q = P_1$ vs. $Q = P_2$. For $T \geq 0$, define a region

$$A_n(T) = \left\{ x^n : \frac{P_1(x_1, x_2, \ldots, x_n)}{P_2(x_1, x_2, \ldots, x_n)} > T \right\}.$$

Let

$$\alpha^* = P_1^n(A_n^c(T)), \quad \beta^* = P_2^n(A_n(T))$$

be the corresponding probabilities of error corresponding to decision region A_n. Let B_n be any other decision region with associated probabilities of error α and β. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Proof of the Neyman-Pearson Lemma

- Let $A = A_n(T)$ be the region defined in the statement.

  Let $B \subseteq \mathcal{X}^n$ be any other acceptance region.

  Let ϕ_A and ϕ_B be the indicator functions of A and B, respectively.

  Claim: For all $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$,

  $$(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) \geq 0.$$

  Suppose, first, that $\boldsymbol{x} \in A$.

  Then we have:

  - If $\boldsymbol{x} \in B$, then $(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) = 0$.
  - If $\boldsymbol{x} \notin B$, then $(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) = P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x}) \geq 0$.

  Suppose, next, that $\boldsymbol{x} \notin A$.

  Then we have:

  - If $\boldsymbol{x} \notin B$, then $(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) = 0$.
  - If $\boldsymbol{x} \in B$, then $(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) = -(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) \geq 0$.

Proof of the Neyman-Pearson Lemma (Cont'd)

- We showed that

$$(\phi_A(\boldsymbol{x}) - \phi_B(\boldsymbol{x}))(P_1(\boldsymbol{x}) - TP_2(\boldsymbol{x})) \geq 0.$$

Multiplying out and summing this over the entire space, we obtain

$$\begin{array}{rcl} 0 & \leq & \sum(\phi_A P_1 - T\phi_A P_2 - P_1\phi_B + TP_2\phi_B) \\ & = & \sum_A(P_1 - TP_2) - \sum_B(P_1 - TP_2) \\ & = & (1-\alpha^*) - T\beta^* - (1-\alpha) + T\beta \\ & = & T(\beta - \beta^*) - (\alpha^* - \alpha). \end{array}$$

Since $T \geq 0$, we have proved the theorem.

Likelihood Ratio Test

- The Neyman-Pearson Lemma indicates that the optimum test for two hypotheses is of the form

$$\frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} > T.$$

- This is the **likelihood ratio test**.
- The ratio is called the **likelihood ratio**.

Example

- Suppose we want to test between two Gaussian distributions,
  - $f_1 = \mathcal{N}(1, \sigma^2)$;
  - $f_2 = \mathcal{N}(-1, \sigma^2)$.

  The likelihood ratio becomes

$$\begin{array}{rcl}\frac{f_1(X_1, X_2, \ldots, X_n)}{f_2(X_1, X_2, \ldots, X_n)} & = & \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-1)^2}{2\sigma^2}}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i+1)^2}{2\sigma^2}}} \\ & = & e^{+\frac{2\sum_{i=1}^n X_i}{\sigma^2}} \\ & = & e^{+\frac{2n\overline{X}_n}{\sigma^2}}.\end{array}$$

  Hence, the likelihood ratio test consists of comparing the sample mean $\overline{X}_n$ with a threshold.

  If we want the two probabilities of error to be equal, we set $T = 1$.

Log-Likelihood Ratio Test in terms of Distances

- We can rewrite the log-likelihood ratio as

$$\begin{aligned} L(X_1, X_2, \ldots, X_n) &= \log \tfrac{P_1(X_1,X_2,\ldots,X_n)}{P_2(X_1,X_2,\ldots,X_n)} \\ &= \textstyle\sum_{i=1}^n \log \tfrac{P_1(X_i)}{P_2(X_i)} \\ &= \textstyle\sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \tfrac{P_1(a)}{P_2(a)} \\ &= \textstyle\sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \tfrac{P_1(a)}{P_2(a)} \tfrac{P_{X^n}(a)}{P_{X^n}(a)} \\ &= \textstyle\sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \tfrac{P_{X^n}(a)}{P_2(a)} \\ &\qquad - \textstyle\sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \tfrac{P_{X^n}(a)}{P_1(a)} \\ &= nD(P_{X^n}\|P_2) - nD(P_{X^n}\|P_1). \end{aligned}$$

  Hence, the likelihood ratio test is equivalent to

$$D(P_{X^n}\|P_2) - D(P_{X^n}\|P_1) > \frac{1}{n}\log T.$$

Informal Argument for Choosing Threshold

- We offer some informal arguments based on Sanov's Theorem to show how to choose the threshold to obtain different probabilities of error.
- Let B denote the set on which hypothesis 1 is accepted.
- The probability of error of the first kind is $\alpha_n = P_1^n(P_{X^n} \in B^c)$.
- The set B^c is convex.
- By Sanov's Theorem, the probability of error is determined essentially by the relative entropy of the closest member of B^c to P_1.
- Therefore,
$$\alpha_n \doteq 2^{-nD(P_1^* \| P_1)},$$
where P_1^* is the closest element of B^c to distribution P_1.
- Similarly,
$$\beta_n \doteq 2^{-nD(P_2^* \| P_2)},$$
where P_2^* is the closest element in B to the distribution P_2.

Informal Argument for Choosing Threshold (Cont'd)

- To obtain the type in B that is closest to P_2, we minimize $D(P\|P_2)$ subject to the constraint $D(P\|P_2) - D(P\|P_1) \geq \frac{1}{n}\log T$.
- Setting up the minimization using Lagrange multipliers, we have

$$J(P) = \sum P(x)\log\frac{P(x)}{P_2(x)} + \lambda\sum P(x)\log\frac{P_1(x)}{P_2(x)} + \nu\sum P(x).$$

- Differentiating with respect to $P(x)$ and setting to 0, we have

$$\log\frac{P(x)}{P_2(x)} + 1 + \lambda\log\frac{P_1(x)}{P_2(x)} + \nu = 0.$$

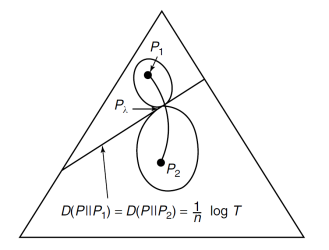
- Solving this set of equations, we obtain the minimizing P

$$P_2^* = P_{\lambda^*} = \frac{P_1^\lambda(x)P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}}P_1^\lambda(a)P_2^{1-\lambda}(a)},$$

where λ is chosen so that $D(P_{\lambda^*}\|P_1) - D(P_{\lambda^*}\|P_2) = \frac{1}{n}\log T$.

Informal Argument for Choosing Threshold (Conclusion)

- From the symmetry of the expression giving P_2^*:
  - $P_1^* = P_2^*$;
  - The probabilities of error behave exponentially with exponents given by the relative entropies $D(P^* \| P_1)$ and $D(P^* \| P_2)$.
- Moreover, we get:
  - $P_\lambda \xrightarrow{\lambda \to 1} P_1$;
  - $P_\lambda \xrightarrow{\lambda \to 0} P_2$.
- The curve that P_λ traces out as λ varies is a geodesic in the simplex.
- P_λ is a normalized convex combination, where the combination is in the exponent.



Subsection 8

Chernoff-Stein Lemma

AEP for Relative Entropy

Theorem (AEP for Relative Entropy)

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to $P_1(x)$, and let $P_2(x)$ be any other distribution on X. Then

$$\frac{1}{n} \log \frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} \to D(P_1 \| P_2) \text{ in probability.}$$

- We apply the Weak Law of Large Numbers.

$$\begin{array}{rcl} \frac{1}{n} \log \frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} & = & \frac{1}{n} \log \frac{\prod_{i=1}^n P_1(X_i)}{\prod_{i=1}^n P_2(X_i)} \\ & = & \frac{1}{n} \sum_{i=1}^n \log \frac{P_1(X_i)}{P_2(X_i)} \\ & \to & EP_1 \log \frac{P_1(X)}{P_2(X)} \text{ in probability} \\ & = & D(P_1 \| P_2). \end{array}$$

Relative Entropy Typical Sets

Definition

For a fixed n and $\epsilon > 0$, a sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ is said to be **relative entropy typical** if and only if

$$D(P_1 \| P_2) - \epsilon \leq \frac{1}{n} \log \frac{P_1(x_1, x_2, \ldots, x_n)}{P_2(x_1, x_2, \ldots, x_n)} \leq D(P_1 \| P_2) + \epsilon.$$

The set of relative entropy typical sequences is called the **relative entropy typical set** $A_\epsilon^{(n)}(P_1 \| P_2)$.

Properties of Relative Entropy Typical Sets

Theorem

1. For $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}(P_1 \| P_2)$,

$$\begin{aligned} P_1(x_1, x_2, \ldots, x_n) 2^{-n(D(P_1\|P_2)+\epsilon)} &\leq P_2(x_1, x_2, \ldots, x_n) \\ &\leq P_1(x_1, x_2, \ldots, x_n) 2^{-n(D(P_1\|P_2)-\epsilon)}. \end{aligned}$$

2. $P_1(A_\epsilon^{(n)}(P_1\|P_2)) > 1 - \epsilon$, for n sufficiently large.
3. $P_2(A_\epsilon^{(n)}(P_1\|P_2)) < 2^{-n(D(P_1\|P_2)-\epsilon)}$.
4. $P_2(A_\epsilon^{(n)}(P_1\|P_2)) > (1-\epsilon) 2^{-n(D(P_1\|P_2)+\epsilon)}$, for n sufficiently large.

1. Property 1 follows directly from the definition of the relative entropy typical set.
2. Property 2 follows from the preceding theorem.

Property 3

3. Write

$$\begin{aligned}
&P_2(A_\epsilon^{(n)}(P_1\|P_2)) \\
&= \textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_2(x_1,x_2,\ldots,x_n) \\
&\leq \textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_1(x_1,x_2,\ldots,x_n)2^{-n(D(P_1\|P_2)-\epsilon)} \\
&\text{(by Property 1)} \\
&= 2^{-n(D(P_1\|P_2)-\epsilon)} \textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_1(x_1,x_2,\ldots,x_n) \\
&= 2^{-n(D(P_1\|P_2)-\epsilon)} P_1(A_\epsilon^{(n)}(P_1\|P_2)) \\
&\leq 2^{-n(D(P_1\|P_2)-\epsilon)}. \\
&\text{(the probability of any set under } P_1 \text{ is less than 1)}
\end{aligned}$$

Property 4

4. To prove the lower bound on the probability of the relative entropy typical set, we use a parallel argument with a lower bound on the probability.

$$\begin{aligned}
&P_2(A_\epsilon^{(n)}(P_1\|P_2))\\
&= \textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_2(x_1,x_2,\ldots,x_n)\\
&\geq \textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_1(x_1,x_2,\ldots,x_n)2^{-n(D(P_1\|P_2)+\epsilon)}\\
&= 2^{-n(D(P_1\|P_2)+\epsilon)}\textstyle\sum_{x^n\in A_\epsilon^{(n)}(P_1\|P_2)} P_1(x_1,x_2,\ldots,x_n)\\
&= 2^{-n(D(P_1\|P_2)+\epsilon)}P_1(A_\epsilon^{(n)}(P_1\|P_2))\\
&\geq (1-\epsilon)2^{-n(D(P_1\|P_2)+\epsilon)}.
\end{aligned}$$

(by Property 2 of $A_\epsilon^{(n)}(P_1\|P_2)$)

Sets of High Probability and Typical Sets

Lemma

Let $B_n \subseteq \mathcal{X}^n$ be any set of sequences $x_1, x_2, \ldots, x_n$, such that $P_1(B_n) > 1 - \epsilon$. Let P_2 be any other distribution such that $D(P_1 \| P_2) < \infty$. Then $P_2(B_n) > (1 - 2\epsilon) 2^{-n(D(P_1\|P_2)+\epsilon)}$.

- For simplicity, we will denote $A_\epsilon^{(n)}(P_1 \| P_2)$ by A_n.

  By the preceding theorem, $P(A_n) > 1 - \epsilon$.

  By hypothesis, $P_1(B_n) > 1 - \epsilon$.

  By the union bound, $P_1(A_n^c \cup B_n^c) < 2\epsilon$.

  Equivalently, $P_1(A_n \cap B_n) > 1 - 2\epsilon$.

Sets of High Probability and Typical Sets

- Now we calculate

$$\begin{array}{rcl} P_2(B_n) & \geq & P_2(A_n \cap B_n) \\ & = & \sum_{x^n \in A_n \cap B_n} P_2(x^n) \\ & \geq & \sum_{x^n \in A_n \cap B_n} P_1(x^n) 2^{-n(D(P_1 \| P_2)+\epsilon)} \\ & & \text{(properties of typical sequences)} \\ & = & 2^{-n(D(P_1 \| P_2)+\epsilon)} \sum_{x^n \in A_n \cap B_n} P_1(x^n) \\ & = & 2^{-n(D(P_1 \| P_2)+\epsilon)} P_1(A_n \cap B_n) \\ & \geq & 2^{-n(D(P_1 \| P_2)+\epsilon)}(1-2\epsilon). \\ & & \text{(by the union bound)} \end{array}$$

The Chernoff-Stein Lemma

- We consider the problem of testing two hypotheses, P_1 vs. P_2.
- We hold one of the probabilities of error fixed.
- We attempt to minimize the other probability of error.
- The relative entropy is the best exponent in probability of error.

Theorem (Chernoff-Stein Lemma)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q$. Consider the hypothesis test between two alternatives, $Q = P_1$ and $Q = P_2$, where $D(P_1 \| P_2) < \infty$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for hypothesis H_1. Let the probabilities of error be $\alpha_n = P_1^n(A_n^c)$, $\beta_n = P_2^n(A_n)$. For $0 < \epsilon < \frac{1}{2}$, define $\beta_n^\epsilon = \min\limits_{\substack{A_n \subseteq \mathcal{X}^n \\ \alpha_n < \epsilon}} \beta_n$. Then

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1 \| P_2).$$

Proof of the Chernoff-Stein Lemma

- We prove this theorem in two parts.
  - In the first part, we exhibit a sequence of sets A_n for which the probability of error β_n goes exponentially to zero as $D(P_1 \| P_2)$.
  - In the second part, we show that no other sequence of sets can have a lower exponent in the probability of error.

  For the first part, we choose as the sets $A_n = A_\epsilon^{(n)}(P_1 \| P_2)$.

  By the preceding theorem, $P_1(A_n^c) < \epsilon$ for n large enough.

  By Property 3 of Relative Entropy Typical Sets,

$$\lim_{n\to\infty} \frac{1}{n} \log P_2(A_n) \leq -(D(P_1 \| P_2) - \epsilon).$$

  Thus, the relative entropy typical set satisfies the bounds of the lemma.

Proof of the Chernoff-Stein Lemma (Cont'd)

- To show that no other sequence of sets can do better, consider any sequence of sets B_n with $P_1(B_n) > 1 - \epsilon$.

  By the preceding lemma, $P_2(B_n) > (1 - 2\epsilon)2^{-n(D(P_1\|P_2)+\epsilon)}$.

  Therefore

$$\begin{aligned} \lim_{n\to\infty} \tfrac{1}{n} \log P_2(B_n) & \\ &> -(D(P_1\|P_2) + \epsilon) + \lim_{n\to\infty} \tfrac{1}{n} \log(1 - 2\epsilon) \\ &= -(D(P_1\|P_2) + \epsilon). \end{aligned}$$

  Thus, no other sequence of sets has a probability of error exponent better than $D(P_1\|P_2)$. Thus, the set sequence $A_n = A_\epsilon^{(n)}(P_1\|P_2)$ is asymptotically optimal in terms of the exponent in the probability.

  Note: The relative entropy typical set is asymptotically optimal, but is not the optimal set for any fixed hypothesis testing problem.

  The optimal set that minimizes the probabilities of error is that given by the Neyman-Pearson lemma.

Subsection 9

Chernoff Information

The Setup

- Instead of treating the two probabilities of error separately, as in the Chernoff-Stein Lemma, we can follow a Bayesian approach, in which we assign prior probabilities to both hypotheses.
- In this case we wish to minimize the overall probability of error given by the weighted sum of the individual probabilities of error.
- The resulting error exponent is the **Chernoff information**.
- Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q$.
- We have two hypotheses:
  - $Q = P_1$ with prior probability π_1;
  - $Q = P_2$ with prior probability π_2.
- The overall probability of error is $P_e^{(n)} = \pi_1 \alpha_n + \pi_2 \beta_n$.
- Let

$$D^* = \lim_{n \to \infty} \left(-\frac{1}{n} \log \min_{A_n \subseteq \mathcal{X}^n} P_e^{(n)} \right).$$

Chernoff's Theorem

Theorem (Chernoff)

The best achievable exponent in the Bayesian probability of error is D^*, where

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

with

$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a \in \mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}$$

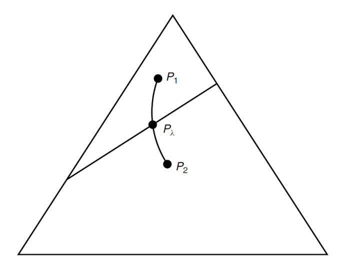
and λ^* the value of λ such that

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

Chernoff's Theorem (Cont'd)

- We have shown that the optimum test is a likelihood ratio test, which can be considered to be of the form

$$D(P_{X^n}\|P_2) - D(P_{X^n}\|P_1) > \frac{1}{n}\log T.$$



- The test divides the probability simplex into regions corresponding to Hypothesis 1 and Hypothesis 2, respectively.

Chernoff's Theorem (Cont'd)

- Let A be the set of types associated with Hypothesis 1.

  By previous work, the closest point in the set A^c to P_1 is on the boundary of A and is of the form given by

  $$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a \in \mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}.$$

  From the discussion in the preceding section, P_λ is:
  - The distribution in A that is closest to P_2;
  - The distribution in A^c that is closest to P_1.

  By Sanov's Theorem, the associated probabilities of error are:
  - $\alpha_n = P_1^n(A^c) \doteq 2^{-nD(P_{\lambda^*} \| P_1)}$;
  - $\beta_n = P_2^n(A) \doteq 2^{-nD(P_{\lambda^*} \| P_2)}$.

Chernoff's Theorem (Cont'd)

- In the Bayesian case, the overall probability of error is the weighted sum of the two probabilities of error,

$$\begin{array}{rcl} P_e & \doteq & \pi_1 2^{-nD(P_{\lambda^*} \| P_1)} + \pi_2 2^{-nD(P_{\lambda^*} \| P_2)} \\ & \doteq & 2^{-n \min \{D(P_{\lambda^*} \| P_1), D(P_{\lambda^*} \| P_2)\}}, \end{array}$$

  since the exponential rate is determined by the worst exponent.

  Now note that:

  - $D(P_\lambda \| P_1)$ increases with λ;
  - $D(P_\lambda \| P_2)$ decreases with λ.

Chernoff's Theorem (Conclusion)

- $D(P_\lambda \| P_1)$ increases with λ and $D(P_\lambda \| P_2)$ decreases with λ.

  So the maximum value of the minimum of

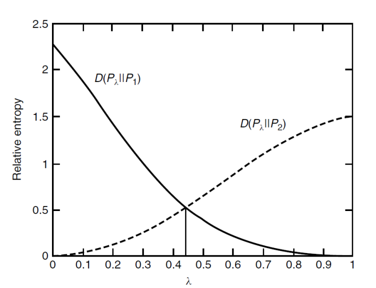
  $$\{D(P_\lambda \| P_1), D(P_\lambda \| P_2)\}$$

  is attained when they are equal.

  Hence, we choose λ so that

  $$D(P_\lambda \| P_1) = D(P_\lambda \| P_2).$$

  This is the highest achievable exponent for the probability of error and is called the **Chernoff information**.



The Chernoff Information

- The definition
$$D^* = D(P_{\lambda^*} \| P_1) = D(P_{\lambda^*} \| P_2)$$
is equivalent to the standard definition of **Chernoff information**,
$$C(P_1, P_2) := - \min_{0 \le \lambda \le 1} \log \left(\sum_x P_1^\lambda(x) P_2^{1-\lambda}(x) \right).$$

Sketching the Derivation of the Chernoff Bound

- The maximum a posteriori probability decision rule minimizes the Bayesian probability of error.

  The decision region A for hypothesis H_1 for this rule is

  $$A = \left\{ \boldsymbol{x} : \frac{\pi_1 P_1(\boldsymbol{x})}{\pi_2 P_2(\boldsymbol{x})} > 1 \right\}.$$

  It comprises the set of outcomes where the a posteriori probability of H_1 is greater than that of H_2.

  The probability of error for this rule is

  $$\begin{array}{rcl} P_e & = & \pi_1 \alpha_n + \pi_2 \beta_n \\ & = & \sum_{A^c} \pi_1 P_1 + \sum_A \pi_2 P_2 \\ & = & \sum \min \{\pi_1 P_1, \pi_2 P_2\}. \end{array}$$

Derivation of the Chernoff Bound (Cont'd)

- For any $a, b > 0$, we have, for all $0 \leq \lambda \leq 1$,

$$\min\{a, b\} \leq a^{\lambda} b^{1-\lambda}.$$

Using this to continue the chain, we have

$$\begin{aligned} P_e &= \sum \min\{\pi_1 P_1, \pi_2 P_2\} \\ &\leq \sum (\pi_1 P_1)^{\lambda} (\pi_2 P_2)^{1-\lambda} \\ &\leq \sum P_1^{\lambda} P_2^{1-\lambda}. \end{aligned}$$

Derivation of the Chernoff Bound (Cont'd)

- For a sequence of i.i.d. observations, $P_k(\boldsymbol{x}) = \prod_{i=1}^n P_k(x_i)$, and

$$\begin{array}{rcl} P_e^{(n)} & \leq & \sum \pi_1^\lambda \pi_2^{1-\lambda} \prod_i P_1^\lambda(x_i) P_2^{1-\lambda}(x_i) \\ & = & \pi_1^\lambda \pi_2^{1-\lambda} \prod_i \sum P_1^\lambda(x_i) P_2^{1-\lambda}(x_i) \\ & \leq & \prod_{x_i} \sum P_1^\lambda P_2^{1-\lambda} \\ & & (\pi_1 \leq 1, \pi_2 \leq 1) \\ & = & (\sum_x P_1^\lambda P_2^{1-\lambda})^n. \end{array}$$

Hence, we have

$$\frac{1}{n} \log P_e^{(n)} \leq \log \sum P_1^\lambda(x) P_2^{1-\lambda}(x).$$

Since this is true for all λ, we can take the minimum over $0 \leq \lambda \leq 1$, resulting in the Chernoff information bound.

This also proves that the exponent is no better than $C(P_1, P_2)$.

Achievability of the bound follows by Chernoff's Theorem.

Remarks on the Error Exponent

- The Bayesian error exponent does not depend on the actual value of π_1 and π_2, as long as they are nonzero.
- The optimum decision rule is to choose the hypothesis with the maximum a posteriori probability, which corresponds to the test
$$\frac{\pi_1 P_1(X_1, X_2, \ldots, X_n)}{\pi_2 P_2(X_1, X_2, \ldots, X_n)} \gtrless 1.$$
- Taking the log and dividing by n, this test can be rewritten as
$$\frac{1}{n}\log\frac{\pi_1}{\pi_2} + \frac{1}{n}\sum_i \log\frac{P_1(X_i)}{P_2(X_i)} \gtrless 0,$$
where the second term tends to $D(P_1\|P_2)$ or $-D(P_2\|P_1)$ accordingly as P_1 or P_2 is the true distribution.
- As the first term tends to 0, the effect of the priors washes out.

Example

- Assume the following data:
  - Major league baseball players have a batting average of 260 with a standard deviation of 15;
  - Minor league ballplayers have a batting average of 240 with a standard deviation of 15.
- A group of 100 ballplayers from one of the leagues, chosen at random, are found to have a group batting average greater than 250.
- They are, therefore, judged to be major leaguers.
- We are now told that we are mistaken; they are minor leaguers.
- What conclusion can be drawn about the distribution of batting averages among these 100 players?
- The Conditional Limit Theorem can be used to show that the distribution of batting averages among these players will have a mean of 250 and a standard deviation of 15.

Abstracting from the Example

- Consider an example of testing between two Gaussian distributions, $f_1 = \mathcal{N}(1, \sigma^2)$ and $f_2 = \mathcal{N}(-1, \sigma^2)$, with different means and the same variance.
- We saw that the likelihood ratio test in this case is equivalent to comparing the sample mean with a threshold.
- The Bayes test is

$$\text{“Accept the hypothesis } f = f_1 \text{ if } \frac{1}{n}\sum_{i=1}^{n} X_i > 0\text{”}.$$

- Now assume that we make an error of the first kind (we say that $f = f_1$ when indeed $f = f_2$) in this test.
- What is the conditional distribution of the samples given that we have made an error?

The Conditional Distribution

- Suppose the true distribution is f_2 and the sample type is in the set A.
- Let f^* denote the distribution in A closest to f_2.
- The conditional distribution is close to f^*.
- By symmetry, this corresponds to $\lambda = \frac{1}{2}$ in

$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a \in \mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}.$$

The Conditional Distribution (Cont'd)

- Calculating the distribution, we get

$$\begin{aligned}
f^*(x) &= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x+1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}}{\int\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x+1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}dx} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2+1}{2\sigma^2}}}{\int\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2+1}{2\sigma^2}}dx} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}} \\
&= \mathcal{N}(0,\sigma^2).
\end{aligned}$$

So the conditional distribution is normal with mean 0 and with the same variance as the original distributions.

Markov's Inequality

Theorem (Markov's Inequality)

For any nonnegative random variable X and any $t > 0$, show that

$$\Pr\{X \geq t\} \leq \frac{EX}{t}.$$

- We have

$$\begin{array}{rcl} EX & = & \int_0^\infty xf(x)x \\ & = & \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \\ & \geq & \int_t^\infty xf(x)dx \\ & \geq & t\int_t^\infty f(x)dx \\ & = & t\Pr\{X \geq t\}. \end{array}$$

So $\Pr\{X \geq t\} \leq \frac{EX}{t}$.

The Chernoff Bound

- The **Chernoff bound** is a special version of Markov's inequality.

Lemma (Chernoff Bound)

Let Y be any random variable and let $\psi(s)$ be the moment generating function of Y, $\psi(s) = Ee^{sY}$. Then, for all $s \geq 0$,

$$\Pr(Y \geq a) \leq e^{-sa}\psi(s).$$

Thus,

$$\Pr(Y \geq a) \leq \min_{s \geq 0} e^{-sa}\psi(s).$$

- Apply Markov's inequality to the nonnegative random variable e^{sY}.

$$\Pr(Y \geq a) = \Pr(e^{sY} \geq e^{sa}) \leq \frac{E^{sY}}{e^{sa}} = \psi(s)e^{-sa}.$$

Subsection 10

Fisher Information and the Cramér-Rao Inequality

The Problem of Parameter Estimation

- A standard problem in statistical estimation is to determine the parameters of a distribution from a sample of data drawn from that distribution.

  Example: Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim \mathcal{N}(\theta, 1)$.

  Suppose that we wish to estimate θ from a sample of size n.

  We can use a number of functions of the data to estimate θ.

  For example, we can use the first sample X_1.

  Although the expected value of X_1 is θ, it is clear that we can do better by using more of the data.

  We guess that the best estimate of θ is the sample mean

  $$\overline{X}_n = \frac{1}{n} \sum X_i.$$

  Indeed, it can be shown that $\overline{X}_n$ is the minimum mean-squared-error unbiased estimator.

Estimators

- Let
$$\{f(x;\theta)\}, \quad \theta \in \Theta,$$
denote an indexed family of densities.
  - $f(x;\theta) \geq 0$;
  - $\int f(x;\theta)dx = 1$, for all $\theta \in \Theta$.
- Θ is called the **parameter set**.

Definition

An **estimator** for θ for sample size n is a function $T : \mathcal{X}^n \to \Theta$.

- Since an estimator is meant to approximate the value of the parameter, we need to have some idea of the goodness of the approximation.
- We call the random variable $T - \theta$ the **error** of the estimator.

Bias of an Estimator

Definition

The **bias** of an estimator $T(X_1, X_2, \ldots, X_n)$ for the parameter θ is the expected value of the error of the estimator, i.e., the bias is

$$E_\theta T(x_1, x_2, \ldots, x_n) - \theta.$$

E_θ is the expectation is with respect to the density $f(\cdot; \theta)$.

The estimator is said to be **unbiased** if the bias is zero, for all $\theta \in \Theta$, i.e., the expected value of the estimator is equal to the parameter.

Example

- Let $X_1, X_2, \ldots, X_n$ drawn i.i.d. according to
$$f(x) = \frac{1}{\lambda} e^{-x/\lambda} \geq 0$$
be a sequence of exponentially distributed random variables.
Estimators of λ include X_1 and $\overline{X}_n$.
We have
$$E_\lambda X_1 - \lambda = \int_0^\infty x_1 \frac{1}{\lambda} e^{-\frac{x_1}{\lambda}} dx_1 - \lambda = \lambda - \lambda = 0.$$
So X_1 is an unbiased estimator.
Similarly,
$$\begin{array}{rcl} E_\lambda \overline{X}_n - \lambda & = & \int_0^\infty \cdots \int_0^\infty \frac{1}{n} \sum_{i=1}^n x_i \prod_{i=1}^n \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}} dx_1 \cdots dx_n - \lambda \\ & = & \lambda - \lambda = 0. \end{array}$$
So $\overline{X}_n$ is also an unbiased estimator.

Consistency of an Estimator in Probability

- The bias is the expected value of the error.
- The fact that it is zero does not guarantee that the error is low with high probability.
- We need to look at some loss function of the error.
- The most commonly chosen is the expected square of the error.
- A good estimator should have:
  - A low expected squared error;
  - An error approaching 0 as the sample size goes to infinity.

Definition

An estimator $T(X_1, X_2, \ldots, X_n)$ for θ is said to be **consistent in probability** if

$$T(X_1, X_2, \ldots, X_n) \to \theta \text{ in probability as } n \to \infty.$$

Domination

- Consistency is a desirable asymptotic property.
- But we are also interested in the behavior for small sample sizes.
- We can then rank estimators on the basis of their mean-squared error.

Definition

An estimator $T_1(X_1, X_2, \ldots, X_n)$ is said to **dominate** another estimator $T_2(X_1, X_2, \ldots, X_n)$ if, for all θ,

$$E(T_1(X_1, X_2, \ldots, X_n) - \theta)^2 \leq E(T_2(X_1, X_2, \ldots, X_n) - \theta)^2.$$

- We would like to discover whether there is a best estimator of θ, i.e., one that dominates every other estimator.
- To answer this question, we derive the Cramér-Rao Lower Bound on the mean-squared error of any estimator.

The Score of a Distribution

Definition

The **score** V is a random variable defined by

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

where $X \sim f(x;\theta)$.

- The mean value of the score is

$$\begin{array}{rcl} EV & = & \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx \\ & = & \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ & = & \frac{\partial}{\partial\theta} \int f(x;\theta)dx \\ & = & \frac{\partial}{\partial\theta} 1 = 0. \end{array}$$

Therefore, $EV^2 = \text{var}(V)$ is the variance of the score.

The Fisher Information

Definition

The **Fisher information** $J(\theta)$ is the variance of the score:

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

- Consider n random variables $X_1, X_2, \ldots, X_n$ drawn i.i.d. $\sim f(x;\theta)$.
  We have $f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i;\theta)$.
  The score function is the sum of the individual score functions,

$$\begin{array}{rcl} V(X_1, X_2, \ldots, X_n) & = & \frac{\partial}{\partial\theta} \ln f(X_1, X_2, \ldots, X_n; \theta) \\ & = & \sum_{i=1}^n \frac{\partial}{\partial\theta} \ln f(X_i;\theta) \\ & = & \sum_{i=1}^n V(X_i). \end{array}$$

  The $V(X_i)$ are independent, identically distributed with zero mean.

The n-Sample Fisher Information

- We found $V(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n V(X_i)$.

  Hence, the n-sample Fisher information is

$$\begin{array}{rcl} J_n(\theta) & = & E_\theta \left[\frac{\partial}{\partial\theta} \ln f(X_1, X_2, \ldots, X_n; \theta)\right]^2 \\ & = & E_\theta V^2(X_1, X_2, \ldots, X_n) \\ & = & E_\theta \left(\sum_{i=1}^n V(X_i)\right)^2 \\ & = & \sum_{i=1}^n E_\theta V^2(X_i) \\ & = & nJ(\theta). \end{array}$$

  Consequently, the Fisher information for n i.i.d. samples is n times the individual Fisher information.

The Cramér-Rao Inequality

Theorem (Cramér-Rao Inequality)

The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:

$$\mathrm{var}(T) \geq \frac{1}{J(\theta)}.$$

- Let V be the score function and T the estimator.
  By the Cauchy-Schwarz inequality,
  $$(E_\theta[(V - E_\theta V)(T - E_\theta T)])^2 \leq E_\theta(V - E_\theta V)^2 E_\theta(T - E_\theta T)^2.$$
  Since T is unbiased, $E_\theta T = \theta$, for all θ.
  We also know that $E_\theta V = 0$. Hence
  $$\begin{array}{rcl} E_\theta(V - E_\theta V)(T - E_\theta T) & = & E_\theta((V - 0)(T - \theta)) \\ & = & E_\theta(VT) - \theta E_\theta V = E_\theta(VT). \end{array}$$

The Cramér-Rao Inequality (Cont'd)

- By definition, $\mathrm{var}(V) = J(\theta)$.

  Substituting these conditions in, we have
  $$[E_\theta(VT)]^2 \leq J(\theta)\mathrm{var}(T).$$
  Now we obtain
  $$\begin{array}{rcl} E_\theta(VT) & = & \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} T(x) f(x;\theta) dx \\ & = & \int \frac{\partial}{\partial\theta} f(x;\theta) T(x) dx \\ & = & \frac{\partial}{\partial\theta} \int f(x;\theta) T(x) dx \\ & & \text{(Bounded Convergence Theorem for nice } f(x;\theta)) \\ & = & \frac{\partial}{\partial\theta} E_\theta T \\ & = & \frac{\partial}{\partial\theta} \theta = 1. \end{array}$$
  So we get $1 \leq J(\theta)\mathrm{var}(T)$, i.e., $\mathrm{var}(T) \geq \frac{1}{J(\theta)}$.

Example

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known.

  We have

  $$J(\theta) = n \int_{-\infty}^{\infty} \left(\frac{x-\theta}{\sigma^2}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} dx = \frac{n}{\sigma^2}.$$

  Let $T(X_1, X_2, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum X_i$. Then

  $$E_\theta(\overline{X}_n - \theta)^2 = \frac{\sigma^2}{n} = \frac{1}{J(\theta)}.$$

  Thus, $\overline{X}_n$ is the minimum variance unbiased estimator of θ, since it achieves the Cramér-Rao lower bound.

Efficient Unbiased Estimators

- The Cramér-Rao inequality gives us a lower bound on the variance for all unbiased estimators.
- When this bound is achieved, we call the estimator efficient.

Definition

An unbiased estimator T is said to be **efficient** if it meets the Cramér-Rao bound with equality, i.e., if

$$\mathrm{var}(T) = \frac{1}{J(\theta)}.$$

- The Fisher information is therefore a measure of the amount of "information" about θ that is present in the data.
- It gives a lower bound on the error in estimating θ from the data.
- It is possible that there is no estimator achieving the bound.