# Chapter 11

# *Information Theory and Statistics*

Peng-Hua Wang

Graduate Inst. of Comm. Engineering

National Taipei University

# Chapter Outline

# 11.1 Method of Types

# Definitions

- Let $X_1, X_2, \ldots$ be a sequence of $n$ symbols from an alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_M\}$ where $M = |\mathcal{X}|$ is the number of alphabets.
- $x^n \equiv \mathbf{x}$ is a sequence $x_1, x_2, \ldots x_n$.
- The **type** $P_{\mathbf{x}}$ (or empirical probability distribution) of a sequence $x_1, x_2, \ldots x_n$ is the relative frequency of each symbol of $\mathcal{X}$.

$$P_{\mathbf{x}}(a) = \frac{N(a|\mathbf{x})}{n}$$

for all $a \in \mathcal{X}$ where $N(a|\mathbf{x})$ is the number of times the symbol $a$ occurs in the sequence $\mathbf{x}$.

**Example.** Let $\mathcal{X} = \{a, b, c\}$, $\mathbf{x} = aabca$. Then the type $P_{\mathbf{x}} = P_{aabca}$ is

$$P_{\mathbf{x}}(a) = \frac{3}{5}, \quad P_{\mathbf{x}}(b) = \frac{1}{5}, \quad P_{\mathbf{x}}(c) = \frac{1}{5}, \quad \text{or } P_{\mathbf{x}} = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right)$$

# Definitions

- The type class $T(P)$ is the set of sequences that have the same type.

$$T(P) = \{\mathbf{x} : P_{\mathbf{x}} = P\}.$$

**Example.** Let $\mathcal{X} = \{a, b, c\}$, $\mathbf{x} = aabca$. Then the type $P_{\mathbf{x}} = P_{aabca}$ is

$$P_{\mathbf{x}}(a) = \frac{3}{5}, \quad P_{\mathbf{x}}(b) = \frac{1}{5}, \quad P_{\mathbf{x}}(c) = \frac{1}{5}.$$

The type class $T(P_{\mathbf{x}})$ is the set of the length-5 sequences that have 3 $a$'s, 1 $b$ and 1 $c$.

$$T(P_{\mathbf{x}}) = \{aaabc, aabca, abcaa, bcaaa, \dots\}.$$

The number of elements in $T(P_{\mathbf{x}})$ is

$$|T(P_{\mathbf{x}})| = \binom{5}{3\ 1\ 1} = \frac{5!}{3!1!1!} = 20.$$

# Definitions

- Let $\mathcal{P}_n$ denote the set of types with denominator $n$. For example, if $\mathcal{X} = \{a, b, c\}$,

$$\mathcal{P}_n = \left\{ \left( \frac{x_1}{n}, \frac{x_2}{n}, \frac{x_3}{n} \right) : x_1 + x_2 + x_3 = n, x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \right\}$$

  where $x_1 = P(a), x_2 = P(b), x_3 = P(c)$.

**Theorem.**

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

**Proof.**

$$\mathcal{P}_n = \left\{ \left( \frac{x_1}{n}, \frac{x_2}{n}, \ldots, \frac{x_M}{n} \right) \right\}$$

where $0 \leq x_k \leq n$. Since there are $n+1$ choices for each $x_k$, the result follows. ■

# Observations

- The number of sequences of length $n$ is $M^n$. (exponential in $n$).
- The number of types of length $n$ is $(n+1)^M$. (polynomial in $n$).
- Therefore, 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.

# Theorem

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

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

where

$$Q^n(\mathbf{x}) = \Pr(\mathbf{x}) = \prod_{i=1}^{n} \Pr(x_i) = \prod_{i=1}^{n} Q(x_i).$$

**Proof.**

$$\begin{aligned} Q^n(\mathbf{x}) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{N(a|\mathbf{x})} \\ &= \prod_{a\in\mathcal{X}} Q(a)^{nP_\mathbf{x}(a)} = \prod_{a\in\mathcal{X}} 2^{nP_\mathbf{x}(a)\log Q(a)} \\ &= 2^{n\sum_{a\in\mathcal{X}} P_\mathbf{x}(a)\log Q(a)} \end{aligned}$$

# Theorem

**Proof. (cont.)** Since

$$\begin{aligned}
&\sum_{a\in\mathcal{X}} P_{\mathbf{x}}(a)\log Q(a)\\
=&\sum_{a\in\mathcal{X}} \left(P_{\mathbf{x}}(a)\log Q(a)+P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a)-P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a)\right)\\
=&-H(P_{\mathbf{x}})-D(P_{\mathbf{x}}||Q),
\end{aligned}$$

we have

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

**Corollary.** If $\mathbf{x}$ is in the type class of $Q$, then

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

**Proof.**

If $\mathbf{x}\in T(Q)$, then $P_{\mathbf{x}}=Q$ and $D(P_{\mathbf{x}}||Q)=0$. $\blacksquare$

# Size of T(P)

Next, we will estimate the size of $|T(P)|$. The exact size of $|T(P)|$ is

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

This value is hard to manipulate. We give a simple bound of $|T(P)|$.

We need the following lemmas.

# Size of T(P)

**Lemma.**

$$\frac{m!}{n!} \geq n^{m-n}$$

**Proof.** For $m \geq n$,we have

$$\begin{aligned}\frac{m!}{n!} &= \frac{1 \times 2 \times \cdots \times m}{1 \times 2 \times \cdots \times n} = (n+1)(n+2) \times \cdots \times m \\ &\geq \underbrace{n \times n \times \ldots n}_{m-n \text{ times}} \\ &= n^{m-n}\end{aligned}$$

For $m < n$,

$$\begin{aligned}\frac{m!}{n!} &= \frac{1 \times 2 \times \cdots \times m}{1 \times 2 \times \cdots \times n} = \frac{1}{(m+1)(m+2) \times \cdots \times n} \\ &\geq \frac{1}{\underbrace{n \times n \times \ldots n}_{n-m}} = \frac{1}{n^{n-m}} = n^{m-n}\end{aligned}$$

■

# Size of T(P)

**Lemma.** The type class $T(P)$ has the the highest probability among all type classes under the probability distribution $P$.

$$P^n(T(P)) \geq P^n(T(\hat{P})) \quad \text{for all } \hat{P} \in \mathcal{P}_n.$$

**Proof.**

$$\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}}$$

$$= \frac{\binom{n}{nP(a_1), nP(a_2), \ldots, nP(a_M)} \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1), n\hat{P}(a_2), \ldots, n\hat{P}(a_M)} \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}}$$

$$= \prod \frac{(n\hat{P}(a))!}{(nP(a))!} P(a)^{n(P(a)-\hat{P}(a))}$$

# Size of T(P)

**Proof. (cont.)**

$$
\begin{aligned}
&\geq \prod_{a\in\mathcal{X}} \left(nP(a)\right)^{n\hat{P}(a)-nP(a)} P(a)^{n(P(a)-\hat{P}(a))} \\
&= \prod_{a\in\mathcal{X}} n^{n\hat{P}(a)-nP(a)} \\
&= n^{n\sum_{a\in\mathcal{X}}\hat{P}(a)-n\sum_{a\in\mathcal{X}}P(a)} \\
&= n^{n-n} = 1 \quad \blacksquare
\end{aligned}
$$

# Size of T(P)

**Theorem.**

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

**Note.** The exact size of $|T(P)|$ is

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

This value is hard to manipulate.

**Proof. (upper bound)**

If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. from $P$, then

$$1 \geq P^n(T(P)) = \sum_{\mathbf{x} \in T(P)} \prod_{a \in \mathcal{X}} P(a)^{nP(a)} = |T(P)| \prod_{a \in \mathcal{X}} 2^{nP(a) \log P(a)}$$

$$= |T(P)| 2^{n \sum_{a \in \mathcal{X}} P(a) \log P(a)} = |T(P)| 2^{-nH(P)}.$$

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

# Size of T(P)

**Proof. (lower bound)**

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

# Probability of type class

**Theorem.** For any $P \in \mathcal{P}_n$ and any distribution $Q$, the probability of the type class $T(P)$ under $Q^n$ satisfies

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

**Proof.**

$$\begin{aligned} Q^n(T(P)) &= \sum (\mathbf{x} \in T(P)) Q^n(\mathbf{x}) \\ &= \sum (\mathbf{x} \in T(P)) 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}||Q))} \\ &= |T(P)| 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}||Q))} \end{aligned}$$

Since

$$\frac{1}{(n+1)^M} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)},$$

we have

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

# Summary

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

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

$$\frac{1}{n}\log|T(P)| \to H(P) \quad \text{as } n \to \infty.$$

$$-\frac{1}{n}\log Q^n(T(P)) \to D(P||Q) \quad \text{as } n \to \infty.$$

- If $X_i \sim Q$, the probability of sequences with type $P \neq Q$ approaches 0 as $n \to \infty$. $\Rightarrow$ Typical sequences are $T(Q)$.

# 11.2 Law of Large Numbers

# Typical Sequences

- Given $\epsilon > 0$, the typical $T_Q^\epsilon$ for the distribution $Q^n$ is defined as

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

- The probability that $\mathbf{x}$ is nontypical is

$$\begin{aligned}
1 - Q^n(T_Q^\epsilon) &= \sum_{P:D(P||Q)>\epsilon} Q^n(T(P)) \\
&\leq \sum_{P:D(P||Q)>\epsilon} 2^{-nD(P||Q)} \\
&\leq \sum_{P:D(P||Q)>\epsilon} 2^{-n\epsilon} \\
&\leq \sum_{P\in Q^n} 2^{-n\epsilon} = (n+1)^M 2^{-n\epsilon} \\
&= 2^{-n(\epsilon - M\frac{\ln(n+1)}{n})}
\end{aligned}$$

# Theorem

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

$$\Pr(D(P_{\mathbf{x}}||P) > \epsilon) \leq 2^{-n(\epsilon - M\frac{\ln(n+1)}{n})}.$$

# 11.3 Universal Source Coding

# Introduction

- An iid source with a known distribution $p(x)$ can be compressed to its entropy $H(X)$. by Huffman coding.
- Wrong code for incorrect distribution $q(x)$, a penalty of $D(p||q)$ bits is incurred.
- Is there a universal code of rate $R$ that is sufficient to compress every iid source with entropy $H(X) < R$?

# Concept

- There are $2^{nH(P)}$ sequences of type $P$.
- There are no more than $(n+1)^{|\mathcal{X}|}$ (polynomial) types.
- There are no more than $(n+1)^{|\mathcal{X}|}2^{nH(P)}$ sequences to describe.
- If $H(P) < R$ there are no more than $(n+1)^{|\mathcal{X}|}2^{nR}$ sequences to describe. Need $nR$ bits as $n \to \infty$.

# 11.4 Large Deviation Theory

# Large Deviation Theory

- If $X_i$ is i.i.d. Bernoulli with $P(X_i = 1) = \frac{1}{3}$, what is the probability that $\frac{1}{n}\sum_{i=1}^{n} X_i$ is near $\frac{1}{3}$? This is a small deviation.
  - Deviation means "deviation from the expected outcome".
  - The probability is near 1.
- What is the probability that $\frac{1}{n}\sum_{i=1}^{n} X_i$ is greater than $\frac{3}{4}$? This is a large deviation.
  - The probability is exponentially small.
  - We might estimate the exponent using central limit theory, but this is a poor approximation for more than a few standard deviation.
  - We note that $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{3}{4}$ is equivalent to $P_{\mathbf{x}} = \left(\frac{1}{4}, \frac{3}{4}\right)$. Thus, the probability is approximated to

$$2^{-nD(P_{\mathbf{x}}||Q)} = 2^{-nD\left(\left(\frac{1}{4},\frac{3}{4}\right)||\left(\frac{1}{3},\frac{2}{3}\right)\right)}$$

# Definition

- Let $E$ be a subset of the set of probability mass functions. We write (with a slight abuse of notation)

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

  - Why do we say this is a slight abuse of notation? The reason is that $Q^n(\cdot)$ in its original meaning represents the probability of a set of sequences. But now we borrow this notion to represent a set of probability mass functions.
  - For example, let $|\mathcal{X}| = 2$ and $E_1$ be the probability mass functions with mean $= -1$. Then $E_1 = \varnothing$.
  - For example, let $|\mathcal{X}| = 2$ and $E_2$ be the probability mass functions with mean $= \sqrt{2}/2$. Then $E_2 \cap \mathcal{P}_n = \varnothing$. (why?)

# Definition

- If $E$ contains a relative entropy neighborhood of $Q$, then by the weak law of large numbers $Q^n(E) \to 1$. Specifically, if $P_{\mathbf{x}} \in E$, then

$$\begin{aligned} Q^n(E) &\geq Q^n(T(P_{\mathbf{x}})) = P(D(P_{\mathbf{x}}||Q) < \epsilon) \\ &\geq 1 - 2^{-n(\epsilon - |\mathcal{X}|\frac{\ln(n+1)}{n})} \to 1 \end{aligned}$$

- Otherwise, $Q^n(E) \to 0$ exponential fast. We will use the method of types to calculate the exponent (rate function.)

# Example

By observation we find that the sample average of $g(X)$ is greater than or equal to $\alpha$. This event is equivalent to the event $P_{\mathbf{X}} \in E \cap \mathcal{P}_n$, where

$$E = \left\{ P : \sum_{a \in \mathcal{X}} g(a) P(a) \geq \alpha \right\}.$$

Because

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i) \geq \alpha \Leftrightarrow \sum_{a \in \mathcal{X}} P_X(a) g(a) \geq \alpha \Leftrightarrow P_{\mathbf{X}} \in E \cap \mathcal{P}_n$$

Thus,

$$\Pr\left( \frac{1}{n} \sum_{i=1}^{n} g(X_i) \geq \alpha \right) = Q^n(E \cap \mathcal{P}_n) = Q^n(E)$$

# Sanov's Theorem

**Theorem.** Let $X_1, X_2, \ldots X_n$ be iid $\sim Q(x)$. 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

$$P^* = \text{argmin}_{P \in E} D(P||Q)$$

is the distribution in $E$ that is closet to $Q$ in relative entropy.

If, in addition, $E \cap \mathcal{P}_n \neq \varnothing$ for all $n \geq n_0$ for some $n_0$, then

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

# Proof of Upper Bound

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

# Proof of Lower Bound

Since $E \cap \mathcal{P}_n \neq \varnothing$ for all $n \geq n_0$, we can find a sequence of $P_n \in E \cap \mathcal{P}_n$ such that $D(P_n||Q) \to D(P^*||Q))$, and

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

Accordingly, $D(P^*||Q)$ is the large deviation rate function.

# Example 1

Suppose that we toss a fair die $n$ times, what is the probability that the average of the throws is greater than or equal to 4 ?

From Sanov's theorem, the large deviation rate function is $D(P^*||Q)$ where $P^*$ minimizes $D(P||Q)$ over all distribution $P$ that satisfy

$$\sum_{i=1}^{6} ip_i \geq 4, \quad \sum_{i=1}^{6} p_i = 1$$

By Lagrange multipliers, we construct the cost function

$$J = D(P||Q) + \lambda \sum_{i=1}^{6} ip_i + \mu \sum_{i=1}^{6} p_i$$

$$= \sum_{i=1}^{6} p_i \ln \frac{p_i}{q_i} + \lambda \sum_{i=1}^{6} p_i + \mu \sum_{i=1}^{6} p_i$$

# Example 1

Let

$$\frac{\partial J}{\partial p_i} = 0 \Rightarrow \ln(6p_i) + 1 + i\lambda + \mu = 0 \Rightarrow p_i = \frac{e^{-1-\mu}}{6} e^{-i\lambda}.$$

Substituting them for the constraints,

$$\sum_{i=1}^{6} ip_i = \frac{e^{-1-\mu}}{6} \sum_{i=1}^{6} ie^{-i\lambda} = 4$$

$$\sum_{i=1}^{6} p_i = \frac{e^{-1-\mu}}{6} \sum_{i=1}^{6} e^{-i\lambda} = 1$$

We can solve numerically $e^{-\lambda} = 1.190804264$. And

$$P^* = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468).$$

# Example 1

The probability that the average of 10000 throws is grater than or equal to 4 is about

$$2^{-nD(P^*||Q)} \approx 2^{-624}$$