Information Theory and Statistics

The Method of Types - Definitions

- The **type** of a sequence $\mathbf{x} = x_1, \ldots, x_n$, denoted by $P_{\mathbf{x}}$, is the relative frequencies in $\mathbf{x}$ of the symbols in $\mathcal{X}$. I.e.

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

where $N(a|\mathbf{x})$ is the number of times a occurs in $\mathbf{x}$.

- $\mathcal{P}_n$ denotes the set of types with denominator n.
- The **type class** of P, denoted by $T(P)$, is defined by

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

- Let $\mathcal{X} = \{1, 2, 3\}$, $\mathbf{x} = 11223$.

$$P_{\mathbf{x}} =? \quad \mathcal{P}_5 =? \quad T(P_{\mathbf{x}}) =?$$

Bounds on the Number of Types

- **Theorem:** The number of types with denominator n is bounded from above by

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

  In other words, there is only a *polynomial* number of types.

- **Proof:** There are $|\mathcal{X}|$ components and each can take a value from $n+1$ possibilities.

The Probability of a Sequence

- The probability of a sequence $\mathbf{x}$ drawn i.i.d. from $Q(x)$ is

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

Proof:

$$Q^n(\mathbf{x}) = \prod_i Q(x_i) = \prod_a Q(a)^{N(a|\mathbf{x})} = \prod_a Q(a)^{nP_{\mathbf{x}}(a)}$$

$$= \prod_a 2^{nP_{\mathbf{x}}(a)\log Q(a)} = 2^{n\sum_a P_{\mathbf{x}}(a)\log Q(a)}$$

$$= 2^{n\sum_a P_{\mathbf{x}}(a)\log \frac{Q(a)}{P_{\mathbf{x}}(a)}P_{\mathbf{x}}(a)} = 2^{n(-H(P_{\mathbf{x}})-D(P_{\mathbf{x}}||Q))}$$

- What does this say about the MLE of $Q(x)$?

The Size of A Type Class

- The size of the type class $T(P)$ of a given type $P \in \mathcal{P}_n$ is bounded by

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

Proof:

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

$$1 = \sum_{P' \in \mathcal{P}_n} P^n(T(P')) \leq |\mathcal{P}_n| P^n(T(P)) = |\mathcal{P}_n| |T(P)| 2^{-nH(P)}$$

Note that $P^n(T(P)) \geq P^n(T(P'))$.

- See Example 12.1.3 for the binary alphabet case.

The Probability of A Type Class

- The probability of the type class $T(P)$ of a given type $P \in \mathcal{P}_n$ under $Q(x)$ is bounded by

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

Proof: We have

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

The result is proved by applying the bounds on $|T(P)|$.

Summary for the Method of Types

- We can summarize the basic theorems as follows
  - $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$
  - $Q^n(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}||Q))}$
  - $|T(P)| \doteq 2^{nH(P)}$
  - $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$
- While the number of sequences is exponential in n, the number of types is polynomial in n.
- The probability of any sequence is exponentially small. In fact, the probability of any type class is exponentially small except for the type class of the true distribution.

Typical Sets

- For $\epsilon > 0$, the typical set T_Q^ϵ of sequences $\mathbf{x} = x_1, \ldots, x_n$ drawn i.i.d. from $Q(x)$ is defined by

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

- The probability that a sequence is not in the typical set goes to 0 as $n \to \infty$. Or

$$Pr(\mathbf{x} \in T_Q^\epsilon) \to 1$$

- The strongly typical set is defined by

$$A_\epsilon^{(n)} = \left\{\mathbf{x} : |\frac{1}{n}N(a|\mathbf{x}) - P(a)| < \frac{\epsilon}{|\mathcal{X}|} \ \forall a\right\}$$

Universal Source Coding

- What compression can be achieved if the true distribution $p(x)$ is unknown?
  - If the wrong distribution $q(x)$ is used, the penalty is $D(p||q)$.
  - But almost certainly a sequence is in the typical set, so $D(p||q)$ can be made small.
- Is there a universal code of rate R that suffices to describe every i.i.d. source with entropy $H < R$? Yes!

Universal Source Coding - Definitions

- A block code of rate R for a source $X_1, \ldots, X_n$ with *unknown* distribution Q encodes and decodes a block of n source symbols at a time.

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$$

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n$$

- Probability of error $P_e^{(n)} = Q^n(\{\mathbf{X} : g_n(f_n(\mathbf{X})) \neq \mathbf{X}\})$
- A code of rate R will be called **universal** if
  - g_n and f_n does not depend on Q
  - $P_e^{(n)} \to 0$ as $n \to \infty$ when $R > H(Q)$.

Universal Source Coding Theorem

- There exists a sequence of $(2^{nR}, n)$ universal source codes such that the $P_e^{(n)} \to 0$ for any source distribution Q with $H(Q) < R$.
- The proof is provided in the next slide. It is based on the fact that the number of sequences of a type increases exponentially with the entropy, while there is only a polynomial number of types.

Proof

Define $R_n = R - |\mathcal{X}|\frac{\log(n+1)}{n}$, $A = \{\mathbf{x} : H(P_{\mathbf{x}}) \leq R_n\}$. Then

$$|A| = \sum_{H(P)\leq R_n} |T(P)| \leq \sum_{H(P)\leq R_n} 2^{nH(P)} \leq \sum_{H(P)\leq R_n} 2^{nR_n} \leq 2^{nR}.$$

So the elements in A can be mapped to $\{1, \ldots, 2^{nR}\}$ with no elements sharing the same integer. It follows that

$$\begin{aligned} P_e^{(n)} = 1 - Q^n(A) &= \sum_{P:H(P)>R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{P:H(P)>R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P:H(P)>R_n} D(P||Q)} \longrightarrow 0 \end{aligned}$$

Properties

- For $H(Q) > R$, a sequence is not in A with high probability. With this code, the error probability is close to 1.
- The scheme descried here works for i.i.d. sources. We will see other schemes which works for non-i.i.d. sources as well.
- Universal codes need a longer block length to achieve the same performance as say Huffman codes, which requires detailed distribution, but not encoding and decoding long blocks.

Large Deviation Theory

- What is the probability that the sample average is close to p or $q \neq p$ for samples drawn i.i.d. from Bernoulli(p)?
- More generally, suppose the true distribution is Q. What is the probability that we observe a sequence of a type in a set E which does not contain Q? That is

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

Sanov's Theorem

- Let $X_1, \ldots, X_n$ be drawn i.i.d. from $Q(x)$. Let E be a set of probability distributions. Then

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

where

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

Furthermore, if the set E includes its closure, then

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

- Proof: see text; Examples: dice, coins.

Hypothesis Testing

- One problem in statistics is to decide between two alternative explanations for the observed data. For example
  - Is a new drug effective?
  - Is a coin biased?
- In the simplest hypothesis testing problem, we want to decide between two i.i.d. distributions for *explaining* the data.

Problem

- Let $\mathbf{X} = X_1, \ldots, X_n$ be i.i.d. random variables with distribution $Q(x)$. We consider two hypotheses

$$H_1 : Q = P_1 \text{ and } H_2 : Q = P_2.$$

- Let g be the decision function, where $g(\mathbf{x}) = i$ implies that H_i is accepted. Define the error probabilities

$$\alpha = Pr(g(\mathbf{x}) = 2 | H_1 \text{ is true}) = P_1(A^c)$$
$$\beta = Pr(g(\mathbf{x}) = 1 | H_2 \text{ is true}) = P_2(A),$$

  where A is the set over which g is 1.

- There is a trade-off between minimizing α, β, so we minimize one subject to a constraint on the other.

Neyman-Pearson Lemma

Follow the previous setting. For $T \geq 0$, define a region

$$A(T) = \left\{ \mathbf{x} : \frac{P_1(x_1, \ldots, x_n)}{P_2(x_1, \ldots, x_n)} > T \right\}$$

Let

$$\alpha^* = P_1(A^c(T)), \ \ \beta^* = P_2(A(T)).$$

These are the error probabilities if we use $A(T)$ as the region with $g = 1$. Let B be any other decision region with associated probabilities of error α, β. Then

$$\alpha \leq \alpha^* \Rightarrow \beta \geq \beta^*.$$

See the text for proof.

Optimal Test for Two Hypotheses

- The optimal test, called the likelihood ratio test, is of the form

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

- The log likelihood can be shown to be the difference between the relative entropies of the sample type to each of the two distributions. I.e.

$$L(\mathbf{X}) = \log \frac{P_1(\mathbf{X})}{P_2(\mathbf{X})} = nD(P_{\mathbf{X}}||P_2) - nD(P_{\mathbf{X}}||P_1)$$

  The likelihood ratio test is now equivalent to

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

Stein's Lemma

We now consider the case where we constraint on one error probability (α) and minimize the other (β).

Let $\mathbf{X} = X_1, \ldots, X_n$ be i.i.d. $\sim Q(x)$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for H_1. Define

$$\alpha_n = P_1(A_n^c), \ \ \beta_n = P_2(A_n),$$

and

$$\beta_n^\epsilon = \min_{A_n \subseteq \mathcal{X}^n, \alpha_n < \epsilon} \beta_n.$$

Then

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1||P_2).$$

See the text for proof.

Lempel-Ziv Coding

- A parsing S of a string is a division of the string into phrases. A distinct parsing is a parsing such that no phrases are identical.
- Lempel-Ziv coding scheme:
  - Apply a distinct parsing to a source string into the shortest phrases
  - The prefix must have appeared in the string.
  - Represent these phrases by the position of the prefix (which is also a phrase) and the last source symbol.
- An example helps to illustrate the idea.

Average Length of Lempel-Ziv Coding

- Let $c(n)$ be the number of pharses in the parsing of a binary input string of length n.
  - $c(n)$ depends on the actual string.
- The compressed representation consists of $c(n)$ pairs of prefix location and last symbol of the phrase. Need $\log c(n)$ bits for prefix location and 1 bit for last symbol.
- The average length (in bits per source symbol) of a Lempel-Ziv coding for a length-n string is thus

$$\frac{c(n)(\log c(n) + 1)}{n}.$$

The Number of Phrases

- **Theorem:** Let $c(n)$ be the number of pharses in the parsing of a binary input string of length n. Then
$$c(n) \leq \frac{n}{(1-\epsilon_n)\log n},$$
where $\epsilon_n \longrightarrow 0$ as $n \longrightarrow \infty$.

- **Proof:** Let n_k be the sum of lengths of all distinct phrases of length no greater then k. I.e.
$$n_k = \sum_{j=1}^{k} j2^j = (k-1)2^{k+1} + 2.$$

Continued Proof

- The number of phrases is maximized when the distinct phrases are as short as possible, so

$$c(n_k) \leq \sum_{j=1}^{k} 2^j \leq \frac{n_k}{k-1}.$$

- For any n, there is one k such that

$$n_k \leq n < n_{k+1}, \;\; c(n) \leq \frac{n}{k-1}, \;\; \text{and} \;\; k \leq \log n.$$

Moreover,

$$n \leq (\log n + 2) 2^{k+2} \Rightarrow (k+2) \geq \log \frac{n}{\log n + 2}.$$

Continued Proof

- It follows that

$$k - 1 \geq \log n - \log(\log n + 2) - 3 \geq (1 - \epsilon_n) \log n,$$

where $\epsilon_n = \min\{1, \frac{\log \log n + 4}{\log n}\}$.

- So the number of phrases in a distinct parsing of a sequence of length n is bounded by

$$c(n) \leq \frac{n}{k-1} \leq \frac{n}{(1 - \epsilon_n) \log n}.$$

A Lemma

- Let Z be a non-negative integer-valued random variable with mean μ. Then

$$H(Z) \leq (\mu + 1)\log(\mu + 1) - \mu \log \mu.$$

- The proof follows from the theory of the maximum entropy distribution.

Markov Approximation

- Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with probability $P(x_1, \ldots, x_n)$.
- The kth order Markov approximation to P is defined by

$$Q_k(x^n_{-(k-1)}) = P(x^0_{-(k-1)}) \prod_{j=1}^{n} P(x_j | x_{j-k}^{j-1}),$$

  where $x_i^j = (x_i, \ldots, x_j)$.
- The entropy rate of the Markov approximation converges

$$-\frac{1}{n} \log Q_k(X_1, \ldots, X_n | X^0_{-(k-1)}) \to H(X_j | X_{j-k}^{j-1}).$$

Preliminary for Ziv's

- Let $(y_1, \ldots, y_c)$ be a distinct parsing for a given string $(x_1, \ldots, x_n)$ into c phrases.
- Let ν_i be the index of the start of the ith phrase. Then

$$y_i = x_{\nu_i}^{\nu_{i+1}-1}, \quad s_i = x_{\nu_i - k}^{\nu_i - 1}.$$

  Note that $s_1 = x_{-(k-1)}^{0}$.
- Let c_{ls} be the number of phrases y_i with length l and preceding state $s_i = s \in \mathcal{X}^k$. Then we have

$$\sum_{l,s} c_{ls} = c, \quad \sum_{l,s} l c_{ls} = n.$$

Ziv's Inequality

- For any distinct parsing of $x_1 x_2 \ldots x_n$, we have

$$\log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq \sum_{l,s} c_{ls} \log \frac{1}{c_{ls}}.$$

- **Proof:**

$$\log Q_k(x_1, x_2, \ldots, x_n | s_1) = \log \prod_{i=1}^{c} P(y_i | s_i) = \sum_{i=1}^{c} \log P(y_i | s_i)$$

$$= \sum_{l,s} \sum_{i:|y_i|=l, s_i=s} \log P(y_i | s_i) = \sum_{l,s} c_{ls} \sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} \log P(y_i | s_i)$$

$$\leq \sum_{l,s} c_{ls} \log (\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} P(y_i | s_i)) \leq \sum_{l,s} c_{ls} \log \frac{1}{c_{ls}}.$$

Main Theorem

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with entropy rate $H(\mathcal{X})$. Let $c(n)$ be the number of phrases in a distinct parsing of a sequence of length n sampled from this process. Then, with probability 1,

$$\lim_{n \to \infty} \sup \frac{c(n) \log c(n)}{n} \le H(\mathcal{X}).$$

So the number of bits per symbol is not greater than the entropy rate.

Proof

From Ziv's inequality, we have

$$-\log Q_k(x_1, x_2, \ldots, x_n | s_1) \geq \sum_{l,s} c_{ls} \log c_{ls} = \sum_{ls} c_{ls} \log \frac{c_{ls} c}{c}$$

$$= c \log c - c \sum_{l,s} \pi_{ls} \log \pi_{ls},$$

where $\pi_{ls} = \frac{c_{ls}}{c}$. Define random variables U, V with

$$Pr(U = l, V = s) = \pi_{ls}.$$

Note $EU = \frac{n}{c}$. Now we have

$$-\frac{1}{n} \log Q_k(x_1, x_2, \ldots, x_n | s_1) \geq \frac{c}{n} \log c - \frac{c}{n} H(U, V).$$

Proof Continued

Now

$$H(U,V) \leq H(V) + H(U)$$
$$\leq k + \log\frac{n}{c} + (\frac{n}{c}+1)\log(\frac{c}{n}+1).$$

From Lemma 12.10.1, $c \sim \frac{n}{\log n}$, so

$$\frac{c}{n}H(U,V) \leq \frac{c}{n}k + \frac{c}{n}\log\frac{n}{c} + o(1) \to 0.$$

Therefore,

$$\frac{c}{n}\log c \leq -\frac{1}{n}\log Q_k(x_1, x_2, \ldots, x_n|s_1) + \epsilon$$
$$\Rightarrow \lim_{n\to\infty}\sup\frac{c(n)\log c(n)}{n} \leq H(\mathcal{X}).$$

Asymptotic Optimality of Lempel-Ziv Coding

- Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with entropy rate $H(\mathcal{X})$. Let $l(X_1, \ldots, X_n)$ be the codeword length of the Lempel-Ziv coding associated with $X_1, \ldots, X_n$. Then, with probability 1,

$$\lim_{n \to \infty} \sup \frac{l(X_1, \ldots, X_n)}{n} \leq H(\mathcal{X})$$

- **Proof:** This follows from

$$l(n) = c(n)(\log c(n) + 1)$$