EE 4TM4: Digital Communications II

Information Measures

Definition 1: The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x).$$

We also write $H(p)$ for the above quantity.

Lemma 1: $H(X) \geq 0$.

Proof: $0 \leq p(x) \leq 1$ implies $\log(1/p(x)) \geq 0$. ■

Definition 2: The relative entropy or Kullback Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

Theorem 1: Jensen's inequality: If f is a convex function and X is a random variable, then

$$Ef(X) \geq f(EX). \tag{1}$$

Moreover, if f is strictly convex, then equality in (1) implies that $X = EX$ with probability 1, i.e., X is a constant.

Theorem 2: Let $p(x)$, $q(x)$, $x \in \mathcal{X}$, be two probability mass functions. Then

$$D(p\|q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof: Let $\mathcal{A} = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p\|q) &= -\sum_{x\in\mathcal{A}} p(x)\log\frac{p(x)}{q(x)} \\
&\sum_{x\in\mathcal{A}} p(x)\log\frac{q(x)}{p(x)} \\
&\leq \log\sum_{x\in\mathcal{A}} p(x)\frac{q(x)}{p(x)} \\
&= \log\sum_{x\in\mathcal{A}} q(x) \\
&\leq \log\sum_{x\in\mathcal{X}} q(x) \\
&= \log 1 \\
&= 0,
\end{aligned}$$

where the first inequality follows from Jensens' inequality. Since $\log t$ is a strictly concave function of t, we have the first inequality becomes equality if and only if $q(x)/p(x) = 1$ everywhere, i.e., $p(x) = q(x)$. Hence we have $D(p\|q) = 0$ if and only if $p(x) = q(x)$ for all x. ■

Definition 3: The type P_{x^n} (or empirical probability distribution) of a sequence $x_1, x_2, \cdots, x_n$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$, i.e., $P_{x^n}(a) = N(a|x^n)/n$ for all $a \in \mathcal{X}$, where $N(a|x^n)$ is the number of times the symbol a occurs in the sequence $x^n \in \mathcal{X}^n$.

Definition 4: Let $\mathcal{P}_n$ denote the set of types with denominator n.

Definition 5: If $P \in \mathcal{P}_n$, then the set of sequences of length n and type P is called the type class of P, denoted $T(P)$, i.e., $T(P) = \{x^n \in \mathcal{X}^n : P_{x^n} = P\}$.

Theorem 3: $|\mathcal{P}_n| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} \leq (n+1)^{|\mathcal{X}|}$.

Theorem 4: If $X_1, X_2, \cdots, X_n$ are drawn i.i.d. according to $Q(x)$, then the probability of x^n depends only on its type and is given by

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

Proof:

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

■

Corollary 1: If x^n is in the type class of Q, then

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

Theorem 5: For any type $P \in \mathcal{P}_n$,

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

Proof: Note that the exact size of $T(P)$ is given by

$$|T(P)| = \binom{n}{nP(a_1), nP(a_2), \cdots, nP(a_{|\mathcal{X}|})} = \frac{n!}{(nP(a_1))!(nP(a_2))!\cdots(nP(a_{|\mathcal{X}|}))!}$$

We first prove the upper bound. Since a type class must have probability ≤ 1, we have

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

Thus $|T(P)| \leq 2^{nH(P)}$. Now for 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, i.e., $P^n(T(P)) \geq P^n(T(\hat{P}))$, for all $\hat{P} \in \mathcal{P}_n$. We lower bound the ratio of probabilities,

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(\hat{P}))} &= \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}} \\
&= \frac{\binom{n}{nP(a_1), nP(a_2), \cdots, nP(a_{|\mathcal{X}|})} \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1), n\hat{P}(a_2), \cdots, n\hat{P}(a_{|\mathcal{X}|})} \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}} \\
&= \prod_{a \in \mathcal{X}} \frac{(n\hat{P}(a))!}{(nP(a))!} P(a)^{n(P(a) - \hat{P}(a))}.
\end{aligned}$$

Now using the simple bound (easy to prove by separately considering the cases $m \geq n$ and $m < n$) $\frac{m!}{n!} \geq n^{m-n}$, we obtain

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(\hat{P}))} &\geq \prod_{a \in \mathcal{X}} (nP(a))^{n\hat{P}(a) - nP(a)} P(a)^{n(P(a) - \hat{P}(a))} \\
&= \prod_{a \in \mathcal{X}} n^{n(\hat{P}(a) - P(a))} \\
&= n^{n(\sum_{a \in \mathcal{X}} \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a))} \\
&= n^{n(1-1)} \\
&= 1.
\end{aligned}$$

Hence $P^n(T(P)) \geq P^n(T(\hat{P}))$. The lower bound now follows easily from this result, since

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

■

Theorem 6: 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,

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

Proof: We have

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

Using the bounds on $|T(P)|$, we have

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

■

Theorem 7: Let $X_1, X_2, \cdots, X_n$ be i.id. $\sim Q(x)$. Then

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

Proof:

$$\begin{aligned} Pr\{D(P_{x^n}\|Q) > \epsilon\} &= \sum_{P: D(P\|Q) > \epsilon} Q^n(T(P)) \\ &\leq \sum_{P: D(P\|Q) > \epsilon} 2^{-nD(P\|Q)} \\ &\leq \sum_{P: D(P\|Q) > \epsilon} 2^{-n\epsilon} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \\ &= 2^{-n(\epsilon - |\mathcal{X}|\frac{\log(n+1)}{n})}. \end{aligned}$$

■

Theorem 8: Sanov's theorem: Let $X_1, X_2, \cdots, X_n$ 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 $P^* = \arg\min_{P \in E} D(P\|Q)$ 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: We first prove the upper bound:

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

Note that P^* need not be a member of $\mathcal{P}_n$. We now come to the lower bound, for which we need a “nice” set E, so that for all large n, we can find a distribution in $E \cap \mathcal{P}_n$ which is close to P^*. If we now assume that E is the closure of its interior (thus the interior must be non-empty), then since $\bigcup_n \mathcal{P}_n$ is dense in the set of all distributions, it follows that $E \cap \mathcal{P}_n$ is non-empty for all $n \geq n_0$ for some $n \geq n_0$. We can then find a sequence of distributions P_n such that $P_n \in E \cap \mathcal{P}_n$ and $D(P_n\|Q) \to D(P^*\|Q)$. For each $n \geq n_0$,

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

Consequently,

$$\liminf \frac{1}{n} \log Q^n(E) \geq \liminf(-\frac{|\mathcal{X}| \log(n+1)}{n} - D(P_n\|Q)) = -D(P^*\|Q).$$

Combining this with the upper bound establishes the theorem. ■

We can summarize the basic theorems concerning types in four equations:

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

Theorem 9: Weak Law of Large Numbers: Let $X_1, X_2, \cdots$ be a sequence of independent random variables having a common distribution Q, and let $E[X_i] = \mu$. Then for any $\epsilon > 0$

$$P\Big\{ \Big| \frac{X_1 + X_2 + \cdots + X_n}{n} - \mu \Big| \geq \epsilon \Big\} \to 0 \quad \text{as } n \to \infty$$

Proof: Let $Y_n = \frac{X_1+X_2+\cdots+X_n}{n} - \mu$. Note that $E[Y_n]=0$ and $\text{Var}(Y_n) = \frac{\sigma^2}{n}$, where $\sigma^2 = \text{Var}(X_i)$. Therefore, by Chebyshev's inequality

$$P\{|Y_n| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty$$

■

Remark: This proof shows that the probability goes to zero at least as fast as $\frac{1}{n}$. In fact, it is possible to derive a tighter bound. Without loss of generality, we assume $E[X_i]=0$. Note that

$$P\{|Y_n| \geq \epsilon\} = P\{|Y_n|^4 \geq \epsilon^4\} \leq \frac{E[Y_n^4]}{\epsilon^4}.$$

Since $E(X_1+X_2+\cdots+X_n)^4 = nE(X^4) + 3n(n-1)(E(X^2))^2$, it follows that the probability goes to zero at least as fast as $\frac{1}{n^2}$.

Remark: According to Sanov's theorem, the probability goes to zero exponentially fast with the exponent given by

$$\min_{P:|\sum_x xP(x)-\mu|\geq\epsilon} D(P\|Q).$$

Theorem 10: Central Limit Theorem: Let $X_1, X_2, \cdots$ be a sequence of independent, identically distributed random variables, each with mean μ and variance σ^2. Then the distribution of

$$\frac{X_1+X_2+\cdots+X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is,

$$P\Big\{\frac{X_1+X_2+\cdots+X_n - n\mu}{\sigma\sqrt{n}} \leq a\Big\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx$$

as $n \to \infty$.

Connection between the central limit theorem and the divergence: Let $X_1, X_2, \cdots$ be a sequence of i.i.d. Bernoulli random variables with parameter p, i.e., $P(X_i = 1) = p$ and $P(X_i = 0) = 1-p$. Note that $E[X_i] = p$ and $\text{Var}(X_i) = E[X_i^2] - E[X_i]^2 = p - p^2 = p(1-p)$. According to the central limit theorem,

$$P\Big\{\frac{X_1+X_2+\cdots+X_n - np}{\sqrt{p(1-p)}\sqrt{n}} \in [a,b]\Big\} \to \frac{1}{\sqrt{2\pi}}\int_{a}^{b} e^{-x^2/2}dx.$$

Note that here we are essentially counting the total probability of type classes such that $X_1+X_2+\cdots+X_n \in [np+a\sqrt{p(1-p)}\sqrt{n}, np+b\sqrt{p(1-p)}\sqrt{n}]$, i.e., the empirical $p_{x^n} \in [p+a\sqrt{p(1-p)}/\sqrt{n}, p+b\sqrt{p(1-p)}/\sqrt{n}]$. Now consider the Taylor expansion of $D(q\|p)$ at the neighborhood of p. Note that

$$D(q\|p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}.$$

So

$$\frac{d}{dq}D(q\|p) = \log\frac{q}{p} - \log\frac{1-q}{1-p},$$

which equals 0 when $q=p$, and

$$\frac{d^2}{dq^2}D(q\|p) = \frac{1}{q(1-q)},$$

which equals $\frac{1}{p(1-p)}$ when $q = p$. Therefore, for $q = p + x\sqrt{p(1-p)}/\sqrt{n}$, we have

$$D(q\|p) \approx \frac{1}{2p(1-p)}(x\sqrt{p(1-p)}/\sqrt{n})^2 = \frac{x^2}{2n}.$$

Consequently, the probability of the type class with empirical $p_{x^n} \approx p + x\sqrt{p(1-p)}/\sqrt{n}$ is approximately

$$e^{-n\frac{x^2}{2n}} = e^{-x^2/2},$$

which gives the right exponent in the central limit theorem. Note that in the limit, one can replace the sum of the probabilities of type classes by integral to recover the central limit theorem (the constant factor $1/\sqrt{2\pi}$ requires a more accurate approximation using Stirling's formula). For the non-binary case, one can also establish a connection between divergence and the multi-dimensional central limit theorem.

Theorem 11: $H(X) \leq \log|\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over $\mathcal{X}$.

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over $\mathcal{X}$, and let $p(x)$ be the probability mass function for X. Then

$$D(p\|u) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X).$$

Hence by the non-negativity of relative entropy,

$$0 \leq D(p\|u) = \log|\mathcal{X}| - H(X).$$

This result can also be proved using Lagrangian multiplier. ■

Conditional entropy:

$$H(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x) = -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x).$$

Conditional entropy: residual uncertainty

Joint entropy:

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y).$$

We have $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$. More generally,

$$H(X^n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1,\cdots,X_{n-1}).$$

Mutual information:

$$\begin{aligned} I(X;Y) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

Note that

$$I(X;Y) = D(p(x,y)\|p(x)p(y)).$$

Therefore, $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.

Since $I(X;Y) \geq 0$, it follows that

$$\begin{aligned} H(Y|X) &\leq H(Y) \\ H(X,Y) &\leq H(X) + H(Y). \end{aligned}$$

Moreover, $H(g(X)) \leq H(X)$ with equality if g is one-to-one over the support of X, i.e., the set $\{x \in \mathcal{X} : p(x) > 0\}$. This is because $H(X,g(X)) = H(g(X)|X) + H(X) = H(X)$ and $H(X,g(X)) = H(g(X)) + H(X|g(X)) \geq H(g(X))$.

Theorem 12: Fano's inequality: Suppose we wish to estimate a random variable X with a distribution $p(x)$. We observe a random variable Y which is related to X by the conditional distribution $p(x|y)$. From Y, we calculate a function $g(Y) = \hat{X}$, which is an estimate of X. We wish to bound the probability that $\hat{X} \neq X$. We observe that $X - Y - \hat{X}$ forms a Markov chain. Define the probability of error

$$P_e = P\{\hat{X} \neq X\}.$$

Then

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y),$$

where $H(P_e) = -P_e \log P_e - (1 - P_e)\log(1 - P_e)$. This inequality can be weakened to $1 + P_e \log|\mathcal{X}| \geq H(X|Y)$.

Proof: Let $E = 1$ if $\hat{X} \neq X$ and $E = 0$ if $\hat{X} = X$. Note that

$$H(E,X|Y) = H(X|Y) + H(E|X,Y) = H(X|Y).$$

On the other hand,

$$\begin{aligned} H(E,X|Y) &= H(E|Y) + H(X|E,Y) \\ &\leq H(E) + H(X|E,Y) \\ &= H(P_e) + H(X|E,Y) \\ &= H(P_e) + P(E=0)H(X|Y,E=0) + P(E=1)H(X|Y,E=1) \\ &\leq H(P_e) + (1 - P_e)0 + P_e \log(|\mathcal{X}| - 1). \end{aligned}$$

■

Conditional mutual information:

$$I(X;Y|Z) = \sum_{z \in \mathcal{Z}} p(z) I(X;Y|Z=z) = \sum_{z \in \mathcal{Z}} p(z) \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

Note that

$$I(X;Y|Z) = \sum_{z\in\mathcal{Z}} p(z)D(p(x,y|z)\|p(x|z)p(y|z)) = D(p(x,y|z)\|p(x|z)p(y|z)|p(z)).$$

Also note that

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z). \end{aligned}$$

The conditional mutual information $I(X;Y|Z)$ is nonnegative and is equal to zero if and only if X and Y are conditionally independent given Z, i.e., $X - Z - Y$ form a Markov chain.

Theorem 13: Data processing inequality: If $X - Y - Z$ form a Markov chain, then

$$I(X;Z) \le I(X;Y).$$

Consequently, for any function g, $I(X;g(Y)) \le I(X;Y)$.

Proof: To prove the data processing inequality, we use the chain rule to expand $I(X;Y,Z)$ in two ways as

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) = I(X;Y) \\ &= I(X;Z) + I(X;Y|Z) \ge I(X;Z). \end{aligned}$$

■

Note that unlike entropy, no general inequality relationship exists between the conditional mutual information $I(X;Y|Z)$ and the mutual information $I(X;Y)$. There are, however two important special cases.

- If X and Z are independent, then

$$I(X;Y|Z) \ge I(X;Y).$$

  This is because $I(X;Y|Z) = I(X;Y,Z) \ge I(X;Y)$.

- If $Z - X - Y$ form a Markov chain, then

$$I(X;Y|Z) \le I(X;Y).$$

  This is because $I(X;Y|Z) \le I(Z,X;Y) = I(X;Y)$.

Typical Sequences

Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables. Then by the (weak) law of large numbers, for each $x \in \mathcal{X}$,

$$N(x|x^n) \to p(x) \quad \text{in probability}.$$

Thus, with high probability, the random empirical pmf $N(x|X^n)$ does not deviate much from the true pmf $p(x)$. For $X \sim p(x)$ and $\epsilon \in (0,1)$, define the set of ϵ-typical n-sequences x^n (or the typical set in short) as

$$\mathcal{T}_\epsilon^{(n)}(X) = \{x^n : |N(x|x^n) - p(x)| \le \epsilon p(x) \text{ for all } x \in \mathcal{X}\}.$$

Note that the typical set can be viewed as the union of type classes whose type is close to $p(x)$. Also note that $p(x)=0$ implies $N(x|x^n)=0$ because otherwise $D(p_{x^n}\|p)=\infty$.

Lemma 2: Typical Average Lemma: Let $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$. Then for any nonnegative function $g(x)$ on $\mathcal{X}$,

$$(1-\epsilon)E(g(X)) \le \frac{1}{n}\sum_{i=1}^{n} g(x_i) \le (1+\epsilon)E(g(X)).$$

Typical sequences satisfy the following properties:

1) Let $p(x^n)=\prod_{i=1}^n p_X(x_i)$. Then for each $x^n \in \mathcal{T}_\epsilon(X)$,

$$2^{-n(H(X)+\delta(\epsilon))} \le p(x^n) \le 2^{-n(H(X)-\delta(\epsilon))},$$

where $\delta(\epsilon)=\epsilon H(X)$. This follows by the typical average lemma with $g(x)=-\log p(x)$.

2) The cardinality of the typical set is upper bounded as

$$|\mathcal{T}_\epsilon^{(n)}| \le 2^{n(H(X)+\delta(\epsilon))}.$$

This can be shown by summing the lower bound in property 1 over the typical set.

3) If $X_1, X_2, \cdots$ are i.i.d. with $X_i \sim p_X(x_i)$, then by the LLN,

$$\lim_{n\to\infty} P\{X^n \in \mathcal{T}_\epsilon^{(n)}\} = 1.$$

This result can also be proved using Theorem 7 in Lecture 1. Note that the size of the typical set is in general negligible compared with $|\mathcal{X}|^n$. However, it captures almost all probability.

4) The cardinality of the typical set is lower bounded as

$$|\mathcal{T}_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\delta(\epsilon))}$$

for n sufficiently large. This follows by property 3 and the upper bound in property 1.

Explain these properties from the perspective of method of types.

Jointly Typical Sequences

Let $(X,Y)\sim p_{X,Y}(x,y)$. The set of jointly ϵ-typical n-sequences is defined as

$$T_\epsilon^{(n)}(X,Y) = \left\{(x^n,y^n) : \left|\frac{1}{n}N(x,y|x^n,y^n) - p_{X,Y}(x,y)\right| \le \epsilon p_{X,Y}(x,y) \text{ for all } (x,y)\in\mathcal{X}\times\mathcal{Y}\right\}.$$

Also define the set of conditionally ϵ-typical n sequences as $\mathcal{T}_\epsilon^{(n)}(Y|x^n) = \left\{y^n : (x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)\right\}$. The properties of typical sequences can be extended to jointly typical sequences as follows.

1) Let $(x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)$ and $p(x^n,y^n)=\prod_{i=1}^n p_{X,Y}(x_i,y_i)$. Then
   (a) $x^n\in\mathcal{T}_\epsilon^{(n)}(X)$ and $y^n\in\mathcal{T}_\epsilon^{(n)}(Y)$,
   (b) $2^{-n(1+\epsilon)H(X)} \le p(x^n) \le 2^{-n(1-\epsilon)H(X)}$ and $2^{-n(1+\epsilon)H(Y)} \le p(y^n) \le 2^{-n(1-\epsilon)H(Y)}$,
   (c) $2^{-n(1+\epsilon)H(X|Y)} \le p(x^n|y^n) \le 2^{-n(1-\epsilon)H(X|Y)}$ and $2^{-n(1+\epsilon)H(Y|X)} \le p(y^n|x^n) \le 2^{-n(1-\epsilon)H(Y|X)}$,
   (d) $2^{-n(1+\epsilon)H(X,Y)} \le p(x^n,y^n) \le 2^{-n(1-\epsilon)H(X,Y)}$.
2) $|\mathcal{T}_\epsilon^{(n)}(X,Y)| \le 2^{n(1+\epsilon)H(X,Y)}$.
3) If $p(x^n,y^n)=\prod_{i=1}^n p_{X,Y}(x_i,y_i)$, then $\lim_{n\to\infty} P\{(X^n,Y^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y)\}=1$.

4) $|\mathcal{T}_\epsilon^{(n)}(X,Y)| \geq (1-\epsilon)2^{n(1-\epsilon)H(X,Y)}$ for n sufficiently large.
5) For every $x^n \in \mathcal{X}^n$, we have $|\mathcal{T}_\epsilon^{(n)}(Y|x^n)| \leq 2^{n(1+\epsilon)H(Y|X)}$.
6) Let $X \sim p_X(x)$ and $Y = g(X)$. Let $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$. Then $y^n \in \mathcal{T}_\epsilon^{(n)}(Y|x^n)$ if and only if $y_i = g(x_i)$ for $i \in [1:n]$.

The following property deserves special attention.

Lemma 3: Conditional Typicality Lemma: Let $(X,Y) \sim p(x,y)$. Suppose that $x^n \in \mathcal{T}_{\epsilon'}^{(n)}(X)$ and $Y^n \sim p(y^n|x^n) = \prod_{i=1}^n p_{Y|X}(y_i|x_i)$. Then, for every $\epsilon > \epsilon'$,

$$\lim_{n\to\infty} P\{(x^n, Y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} = 1.$$

The proof of this lemma follows by the LLN.

The conditional typicality lemma implies the following additional property of jointly typical sequences.

5) If $x^n \in \mathcal{T}_{\epsilon'}^{(n)}(X)$ and $\epsilon' < \epsilon$, then for n sufficiently large,

$$|\mathcal{T}_\epsilon^{(n)}(Y|x^n)| \geq (1-\epsilon)2^{n(1-\epsilon)H(X|Y)}.$$

Joint Typicality for a Triple of Random Variables

Let $(X,Y,Z) \sim p(x,y,z)$. The set of jointly ϵ-typical (x^n,y^n,z^n) sequences is defined as

$$\mathcal{T}_\epsilon^{(n)}(X,Y,Z) = \{(x^n,y^n,z^n) : |N(x,y,z|x^n,y^n,z^n) - p(x,y,z)| \leq \epsilon p(x,y,z) \text{ for all } (x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\}.$$

Suppose that $(x^n,y^n,z^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)$ and $p(x^n,y^n,z^n) = \prod_{i=1}^n p_{X,Y,Z}(x_i,y_i,z_i)$. Then

1) $x^n \in \mathcal{T}_\epsilon^{(n)}$ and $(y^n,z^n) \in \mathcal{T}_\epsilon^{(n)}(Y,Z)$,
2) $p(x^n,y^n,z^n) \approx 2^{-nH(X,Y,Z)}$,
3) $p(x^n,y^n|z^n) \approx 2^{-nH(X,Y|Z)}$,
4) $|\mathcal{T}_\epsilon^{(n)}(X|y^n,z^n)| \leq 2^{n(H(X|Y,Z)+\delta(\epsilon))}$, and
5) if $(y^n,z^n) \in \mathcal{T}_{\epsilon'}^{(n)}(Y,Z)$ and $\epsilon' < \epsilon$, then for n sufficiently large, $|\mathcal{T}_\epsilon^{(n)}(X|y^n,z^n)| \geq 2^{n(H(X|Y,Z)-\delta(\epsilon))}$.

The following two-part lemma will be used in many achievability proofs of coding theorem.

Lemma 4: Joint Typicality Lemma: Let $(X,Y,Z) \sim p(x,y,z)$ and $\epsilon' < \epsilon$. Then there exists $\delta(\epsilon) > 0$ that tends to zero as $\epsilon \to 0$ such that the following statements hold:

1) If $(\tilde{x}^n, \tilde{y}^n)$ is a pair of arbitrary sequences and $\tilde{Z}^n \sim \prod_{i=1}^n p_{Z|X}(\tilde{z}_i|\tilde{x}_i)$, then

$$P\{(\tilde{x}^n,\tilde{y}^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} \leq 2^{-n(I(Y;Z|X)-\delta(\epsilon))}.$$

2) If $(x^n,y^n) \in \mathcal{T}_{\epsilon'}^{(n)}$ and $\tilde{Z}^n \sim \prod_{i=1}^n p_{Z|X}(\tilde{z}_i|x_i)$, then for n sufficiently large,

$$P\{(x^n,y^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} \geq 2^{-n(I(Y;Z|X)+\delta(\epsilon))}.$$

Proof: To prove the first statement, consider

$$\begin{aligned}P\{(\tilde{x}^n, \tilde{y}^n, \tilde{Z}^n \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z))\} &= \sum_{\tilde{z}^n \in \mathcal{T}_\epsilon^{(n)}(Z|\tilde{x}^n, \tilde{y}^n)} p(\tilde{z}^n|\tilde{x}^n)\\ &\leq |\mathcal{T}_\epsilon^{(n)}(Z|\tilde{x}^n, \tilde{y}^n)| 2^{-n(H(Z|X)-\epsilon H(Z|X))}\\ &\leq 2^{n(H(Z|X,Y)+\epsilon H(Z|X,Y))} 2^{-n(H(Z|X)-\epsilon H(Z|X))}\\ &= 2^{-n(I(Y;Z|X)-\delta(\epsilon))}.\end{aligned}$$

Similarly, for every n sufficiently large,

$$\begin{aligned}P\{(x^n, y^n, \tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} &\geq |\mathcal{T}_\epsilon^{(n)}(Z|x^n, y^n)| 2^{-n(H(Z|X)+\epsilon H(Z|X))}\\ &\geq (1-\epsilon) 2^{n(H(Z|X,Y)-\epsilon H(Z|X,Y))} 2^{-n(H(Z|X)+\epsilon H(Z|X))}\\ &= 2^{-n(I(Y;Z|X)+\delta(\epsilon))},\end{aligned}$$

which proves the second statement. ■

References

[1] T. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition. Hoboken, NJ: Wiley, 2006.

[2] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2011.