Electrical Engineering 229A Lecture 7 Notes

Daniel Raban

September 16, 2021

1 Types, Typicality Sets, and Entropy Rate

1.1 Types

Let $\mathscr{X}$ be a finite set (called the alphabet). Given a sequence of symbols $x_1^n := (x_1, \dots, x_n)$ taking values in $\mathscr{X}^n$ and $x \in \mathscr{X}$, let $N(x \mid x_1^n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = x\}}$ be the number of times x shows up in x_1^n. Notice that $(\frac{N(x|x_1^n)}{n}, x \in \mathscr{X})$ is a probability distribution on $\mathscr{X}$ (which depends on $\mathscr{X}$).

Definition 1.1. The distribution $P_{x_1^n} = (\frac{N(x|x_1^n)}{n}, x \in \mathscr{X})$ is clled the **type** of x_1^n in information theory and the **empirical distribution** of x_1^n more generally.

A type based on a sample of size n from $\mathscr{X}$ has to be of the form $(\frac{k_x}{n}, x \in \mathscr{X})$ for some integers $0 \leq k_x \leq n$ with $\sum_x k_x = n$. $\mathcal{P}_n$ denotes the set of all types based on samples of size n from $\mathscr{X}$.

Proposition 1.1.

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

So $|\mathcal{P}_n|$ grows only polynomially in n. Contrast this with the total number of sequences of length n, whose size is $|\mathscr{X}|^n$, exponential in n.

1.2 The scale of typicality sets

Definition 1.2. For $p \in \mathcal{P}_n$, the set $T(p) = \{x_1^n : P_{x_1^n} = p\} \subseteq \mathscr{X}^n$ is called the **typicality set** of type p.

Now note that given any probability distribution $(q(x), x \in \mathscr{X})$ and any sequence $x_1^n \in \mathscr{X}^n$, $q^n(x_1^n) = \prod_{i=1}^n q(x_i)$ is determined by $P_{x_1^n}$, the type of x_1^n, because

$$q^n(x_1^n) = \prod_{x \in \mathscr{X}} q(x)^{N(x|x_1^n)}$$

$$
\begin{aligned}
&= \prod_{x \in \mathscr{X}} 2^{n P_{x_1^n}(x) \log q(x)} \\
&= 2^{n \sum_x P_{x_1^n}(x) \log q(x)},
\end{aligned}
$$

which depends on x_1^n only through its type. But also note that

$$
\sum_x P_{x_1^n}(x) \log q(x) = \sum_x P_{x_1^n}(x) \log \frac{q(x)}{P_{x_1^n}(x)} + \sum_x P_{x_1^n}(x) \log P_{x_1^n}(x),
$$

so

$$
q^n(x_1^n) = 2^{-n(H(P_{x_1^n}) + D(p_{x_1^n} || q))}.
$$

This calculation implies the following:

Proposition 1.2. *For any $p \in \mathcal{P}_n$,*

$$
|T(p)| \leq 2^{nH(p)}.
$$

Proof. Take q to be p and consider x_1^n having $P_{x_1^n} = p$. This tells us that for all x_1^n with type $P_{x_1^n} = p$,

$$
p^n(x_1^n) = 2^{-nH(p)}
$$

because $D(p \mid\mid p) = 0$.

But, given $p \in \mathcal{P}_n$,

$$
\begin{aligned}
1 &= \sum_{x_1^n} p^n(x_1^n) \\
&\geq \sum_{x_1^n : P_{x_1^n} = p} p^n(x_1^n) \\
&= \sum_{x_1^n : P_{x_1^n} = p} 2^{-nH(p)} \\
&= |T(p)| 2^{-nH(p)}.
\end{aligned}
$$

□

We can also prove a lower bound:

Proposition 1.3. *For all $p \in \mathcal{P}_n$,*

$$
|T(p)| \geq \frac{2^{nH(p)}}{(n+1)^{|\mathscr{X}|}}.
$$

Proof. This comes from showing that for $p \in \mathcal{P}_n$, $p^n(T(p)) \geq p^n(T(\widehat{p}))$ for all $\widehat{p} \in \mathcal{P}_n$. The left hand side is

$$p^n(T(p)) = \sum_{x_1^n : P_{x_1^n} = p} p^n(x_1^n) = \sum_{x_1^n : P_{x_1^n} = p} 2^{-nH(p)} = |T(p)|2^{-nH(p)},$$

while the right hand side is $|T(\widehat{p})|2^{-n(H(\widehat{p})+D(\widehat{p}||p))}$.

Substituting the exact values of $|T(p)|$ and $|T(\widehat{p})|$ using combinatorics, the left hand side is $\binom{n}{np(x_1),\dots,np(x_d)}2^{-nH(p)}$ (with $\mathscr{X} = \{a_1, \dots, a_d\}$), while the right hand side is $\binom{n}{n\widehat{p}(a_1),\dots,n\widehat{p}(a_d)}2^{-n(H(\widehat{p})+D(\widehat{p}||p))}$. So

$$\frac{p^n(T(p))}{p^n(T(\widehat{p}))} = \frac{n!}{np(a_1)!\cdots np(a_d)!} \frac{2^{n\sum_{i=1}^d p(a_i)\log p(a_i)}}{n!} \frac{n\widehat{p}(a_1)!\cdots n\widehat{p}(a_d)!}{2^{n\sum_{i=1}^d \widehat{p}(a_i)\log \widehat{p}(a_i)}}$$

Now observe that $\frac{m!}{\ell!} \geq \ell^{m-\ell}$ for all ℓ, m.

$$\begin{aligned}
&\geq \frac{\prod_{i=1}^n p(a_i)^{np(a_i)}(np(a_i))^{n\widehat{p}(a_i)}}{\prod_{i=1}^n \widehat{p}(a_i)^{n\widehat{p}(a_i)}(np(a_i))^{np(a_i)}} \\
&= 1.
\end{aligned}$$

Finally, we have

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

□

1.3 ε-typical sets in terms of types

For a probability distribution q on $\mathscr{X}$,

$$A_\varepsilon^{(n)} := \{x_1^n : \frac{1}{n}\sum_{i=1}^n \log q(x_i) - H(q)| < \varepsilon\}.$$

Proposition 1.4.

$$A_\varepsilon^{(n)} = \{x_1^n : |D(P_{x_1^n} \mid\mid q) + H(P_{x_1^n}) - H(q)| < \varepsilon\}.$$

Proof.

$$-\frac{1}{n}\sum_{i=1}^n \log q(x_i) = -\frac{1}{n}\sum_x N(x \mid x_1^n)\log q(x)$$

$$
\begin{aligned}
&= -\sum_x p_{x_1^n}(x)\log q(x) \\
&= D(P_{x_1^n} \,||\, q) + H(P_{x_1^n}).
\end{aligned}
$$

So

$$A_\varepsilon^{(n)} = \{x_1^n : |D(P_{x_1^n} \,||\, q) + H(P_{x_1^n})| < \varepsilon\},$$

as claimed. □

1.4 Stationary sequences and entropy rate

Beyond iid sequences, we consider stationary random sequences.

Definition 1.3. As sequence of random variables $(X_k)_{k=-\infty}^{\infty}$ with $X_k \in \mathscr{X}$ is called **stationary** if

$$\mathbb{P}(X_\ell = x_0, X_{\ell+1} = x_1, \ldots, X_{\ell+L} = x_L) = \mathbb{P}(X_{\ell+m} = x_0, X_{\ell+m+1} = x_1, \ldots, X_{\ell+m+L} = x_L)$$

for all $\ell, m \in \mathbb{Z}$, $L \geq 0$, and $x_0, \ldots, x_L \in \mathscr{X}$.

For a stationary sequence,

$$H(X_2 \mid X_1) \leq H(X_2),$$

but $H(X_2) = H(X_1)$ by stationarity, so

$$H(X_2 \mid X_1) \leq H(X_1).$$

Similarly,

$$H(X_{L+2} \mid X_1, \ldots, X_{L+1}) \leq H(X_{L+1} \mid X_1, \ldots, X_L)$$

because the left hand side equals $H(X_{L+1} \mid X_0, \ldots, X_L)$ by stationarity.

This implies that for a stationary process,

$$\lim_{L\to\infty} H(X_{L+1} \mid X_1, \ldots, X_L)$$

exists and is called the **entropy rate** of the process. In fact, the chain rule says that this equals

$$\lim_{L\to\infty} \frac{1}{L} H(X_1, \ldots, X_L).$$

Definition 1.4. A stationary process is a **stationary Markov chain** if

$$\mathbb{P}(X_{L+1} = x_{L+1} \mid X_1 = x_1, \ldots, X_L = x_L) = \mathbb{P}(X_{L+1} = x_{L+1} \mid X_L = x_L)$$

for all $L \geq 1$ and $x_1, \ldots, x_{L+1}$.

So all that matters is the matrix $[p(j \mid i) : 1 \leq i, j \leq |\mathscr{X}|]$, where the **transition probabilities** $p(j \mid i) = \mathbb{P}(X_2 = j \mid X_1 = i)$. If we let $\pi(i) := \mathbb{P}(X_1 = i)$ for $i \in \mathscr{X}$ in a stationary Markov chain, then

$$\sum_i \pi(i) p(j, i) = \pi(j)$$

for all j. The entropy rate for a stationary markov chain will be $H(X_2 \mid X_1)$ because $H(X_2 \mid X_1, X_0) = H(X_2, X_1)$. So the entropy rate is

$$\sum_i \pi(i) \sum_j p(j \mid i) \log \frac{1}{p(j \mid i)}.$$