ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 3/12/20

Lecture 12: Properties of Letter Typical Sequences

Lecturer: Prof. Ziv Goldfeld

Scriber: Zhengxin Zhang, *Net ID*: zz658

Assistant Editor: Kia Khezeli

In this lecture, we introduce the notion of letter typical sequences and discuss their properties.

Definition 12.1 (Letter Typical Set) *Let $\mathcal{X}$ be a finite alphabet, $P \in \mathcal{P}(\mathcal{X})$ have a PMF p, $n \in \mathbb{N}$ and $\epsilon > 0$. The $\epsilon-$letter typical set of n-lengthed sequences with respect to P with slackness ϵ is*

$$\mathcal{T}_\epsilon^{(n)}(P) := \left\{x^n \in \mathcal{X}^n : |\nu_{x^n}(a) - p(a)| \leq \epsilon p(a),\ \forall a \in \mathcal{X}\right\}$$

where $\nu_{x^n}(a)$ is the empirical frequency of a in $x^n = (x_1, \cdots, x_n)$.

Remark 12.1 (Initial Observation) *There are several comments in order.*

1. *If $p(a) = 0$, then $\nu_{x^n}(a) = 0$ for all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$, meaning that zero probability letters cannot appear in letter typical sequences.*
2. *Suppose $\epsilon < 1$. If $p(a) > 0$, then $\nu_{x^n}(a) > 0$ for all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$ (indeed, if $\nu_{x^n}(a) = 0$ for some x^n, then we have $p(a) \leq \epsilon p(a)$, which is a contradiction). This means that all letters with positive mass must appear at least once in letter typical sequences.*
3. *If $P = \mathsf{Unif}(\mathcal{X})$, then for all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$,*

$$\frac{(1-\epsilon)}{|\mathcal{X}|} \leq \nu_{x^n}(a) \leq \frac{(1+\epsilon)}{|\mathcal{X}|}.$$

It follows that for small enough ϵ, it holds that $\mathcal{T}_\epsilon^{(n)}(P) \subsetneq \mathcal{X}^n$. For example, if $\epsilon < |\mathcal{X}| - 1$, then $\mathcal{T}_\epsilon^{(n)}(P)$ does not contain any sequence comprised of a single letter.

The strength of Letter Typical Sequences lie in the fact that they land themselves well for empirical averages, as shown in the following lemma.

Lemma 12.1 (Typical Averaging Lemma) *Let $g : \mathcal{X} \to \mathbb{R}$ be a nonnegative measurable function such that $\mathbb{E}_P[g(X)] < \infty$, then for all $x^n \in \mathcal{T}_\epsilon^n(P)$, we have*

$$(1-\epsilon)\mathbb{E}_P[g(X)] \leq \frac{1}{n}\sum_{i=1}^n g(x_i) \leq (1+\epsilon)\mathbb{E}_P[g(X)].$$

Proof: For all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$ and $a \in \mathcal{X}$, we have $|\nu_{x^n}(a) - p(a)| \leq \epsilon p(a)$, where p is the PMF of X. It follows that $\frac{1}{n}\sum_{i=1}^n g(x_i) = \sum_{a\in\mathcal{X}} \nu_{x^n}(a) g(a)$. Then

$$\left|\frac{1}{n}\sum_{i=1}^n g(x_i) - \mathbb{E}_P[g(X)]\right| = \left|\sum_{a\in\mathcal{X}} \big(\nu_{x^n}(a) - p(a)\big) g(a)\right| \leq \sum_{a\in\mathcal{X}} \big|\nu_{x^n}(a) - p(a)\big| g(a) \leq \epsilon \sum_{a\in\mathcal{X}} p(a)g(a) = \epsilon \mathbb{E}_P[g],$$

where the first inequality follows from the triangle inequality. ■

Based on this lemma we derive further properties of $\mathcal{T}_\epsilon^{(n)}(P)$. In particular, we provide upper and lower bounds on the cardinality and the probability of $\mathcal{T}_\epsilon^{(n)}(P)$.

Theorem 1 (Properties of Letter Typical Sequences) *Suppose $\mathcal{T}_\epsilon^{(n)}(P)$ is an ϵ-letter typical set. Let $P^{\otimes n}$ be the n-fold product measure induced by P, i.e., $P^{\otimes n}(\{x^n\}) = \prod_{i=1}^n p(x_i)$, for all $x^n \in \mathcal{X}^n$. The following hold:*

1. *For all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$,*
$$2^{-n(1+\epsilon)H(P)} \leq P^{\otimes n}(\{x^n\}) \leq 2^{-n(1-\epsilon)H(P)}, \tag{1}$$
where $H(P)$ is the Shannon entropy of P.

2. *We have*
$$\lim_{n\to\infty} P^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P)\right) = 1. \tag{2}$$

3. *For n sufficiently large,*
$$(1-\epsilon)2^{n(1-\epsilon)H(P)} \leq \left|\mathcal{T}_\epsilon^{(n)}(P)\right| \leq 2^{n(1+\epsilon)H(P)}. \tag{3}$$

Proof: 1. This an immediate consequence of Lemma 12.1. Let $g(x) = -\log(p(x))$. Note that $\mathbb{E}_P[g(X)] = H(P)$, and $\frac{1}{n}\sum_{i=1}^n g(x_i) = \frac{1}{n}\log\left(\frac{1}{P^{\otimes n}(\{x^n\})}\right)$. The lemma then implies

$$(1-\epsilon)H(P) \leq \frac{1}{n}\log\left(\frac{1}{P^{\otimes n}(\{x^n\})}\right) \leq (1+\epsilon)H(P).$$

By taking the exponential of all terms, we get

$$2^{-n(1-\epsilon)H(P)} \geq P^{\otimes n}(\{x^n\}) \geq 2^{-n(1+\epsilon)H(P)}.$$

2. By definition we have

$$P^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P)\right) = P^{\otimes n}\left(\bigcap_{a\in\mathcal{X}} \left\{x^n : |\nu_{x^n}(a) - p(a)| \leq \epsilon p(a)\right\}\right).$$

By the weak law of large number (WLLN), we have that for a set of arbitrary functions $\{f_k(x)\}_{k=1}^K$ (each with finite expectation) and any $\delta > 0$,

$$\lim_{n\to\infty} P^{\otimes n}\left(\bigcap_{k=1}^K \left\{x : \left|\frac{1}{n}\sum_{i=1}^n f_k(x_i) - \mathbb{E}_P\big[f_k(X)\big]\right| \leq \delta\right\}\right) = 1.$$

Now, set $K = |\mathcal{X}|$, and $f_a = \mathbb{1}_{\{a\}}$, for each $a \in \mathcal{X}$. Then, $\mathbb{E}_P[f_a(X)] = p(a)$, and $\frac{1}{n}\sum_{i=1}^n f_a(x_i) = \nu_{x^n}(a)$. The WLLN then implies

$$\lim_{n\to\infty} P^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P)\right) = \lim_{n\to\infty} P^{\otimes n}\left(\bigcap_{a\in\mathcal{X}} \{x^n : |\nu_{x^n}(a) - p(a)| \leq \epsilon p(a)\}\right) = 1. \tag{4}$$

3. Using the fact that $\mathcal{T}_\epsilon^{(n)}(P) \subseteq \mathcal{X}^n$, we get

$$1 = P^{\otimes n}(\mathcal{X}^n) \geq P^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P)\right) = \sum_{x^n\in\mathcal{T}_\epsilon^{(n)}(P)} P^{\otimes n}(\{x^n\}) \geq \left|\mathcal{T}_\epsilon^{(n)}(P)\right|2^{-n(1+\epsilon)H(P)},$$

where the last inequality follows from (1). So, $\left|\mathcal{T}_\epsilon^{(n)}(P)\right| \leq 2^{n(1+\epsilon)H(P)}$.

For the other direction, for n sufficiently large, by (2) we have $1-\epsilon \leq P^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P)\right)$. Then,

$$1-\epsilon \leq \sum_{x^n\in\mathcal{T}_\epsilon^{(n)}(P)} P^{\otimes n}(x) \leq \left|\mathcal{T}_\epsilon^{(n)}(P)\right|2^{-n(1-\epsilon)H(P)},$$

implying that $\left|\mathcal{T}_\epsilon^{(n)}(P)\right| \geq (1-\epsilon)2^{n(1-\epsilon)H(P)}$.

■

Remark 12.2 (Refined Result) *A finite sample bound on the probability of $\mathcal{T}_\epsilon^{(n)}(P)$ can be derived using a refined argument (based on Chernoff bounds). This gives:*

$$1 - \delta_\epsilon(P, n) \leq P^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P)) \leq 1, \tag{5}$$

where $\delta_\epsilon(P, n) := 2|\mathcal{X}|e^{-2n\epsilon^2\mu^2}$ and $\mu := \min_{a\in\mathcal{X}:\ p(a)>0} p(a)$. Note that $\lim_{n\to\infty} \delta_\epsilon(P, n) = 0$, for any fixed $\epsilon > 0$. Using (5), one can show that

$$(1 - \delta_\epsilon(P, n))\, 2^{n(1-\epsilon)H(P)} \leq \left|\mathcal{T}_\epsilon^{(n)}(P)\right| \leq 2^{n(1+\epsilon)H(P)},$$

for all $n \in \mathbb{N}$. This strengthens Properties 1 and 2 in Theorem 1.

Remark 12.3 (Interpretation and Illustration) *As n goes to infinity, the set $\mathcal{T}_\epsilon^{(n)}(P)$ takes a smaller and smaller portion of the whole space $\mathcal{X}^n$, while the probability tends to concentrate on this set, uniformly distributed across its elements. The following diagram is a useful illustration:*

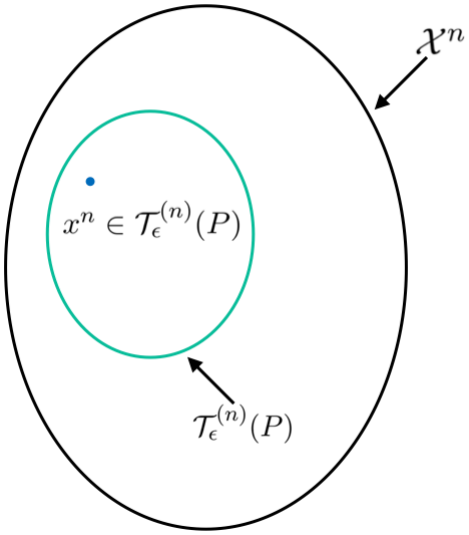

Figure 1: Illustration of letter typical set.

We can further simplify the statement by introducing the following notation. Denote by $a_n \doteq 2^{nb}$, the inequality $2^{n(b-\delta(\epsilon))} \leq a_n \leq 2^{n(b+\delta(\epsilon))}$ if there exists a $\delta(\epsilon) = o(1)$ as ϵ goes to 0. Then, by taking $\delta(\epsilon) = 2\epsilon H(P)$, we can write $|\mathcal{T}_\epsilon^{(n)}(P)| \doteq 2^{nH(P)}$, and $P^{\otimes n}(\{x^n\}) \doteq 2^{-nH(p)}$. The properties of $\mathcal{T}_\epsilon^{(n)}(P)$ given in Theorem 1 can now be understood as follows:

1. *The cardinality of $\mathcal{T}_\epsilon^{(n)}(P) \subsetneqq \mathcal{X}^n$ is much smaller than that of $\mathcal{X}^n$. More precisely, for all $P \neq \mathrm{Unif}(\mathcal{X})$ and sufficiently small ϵ, it holds that*

$$\lim_{n\to\infty} \frac{|\mathcal{T}_\epsilon^{(n)}(P)|}{|\mathcal{X}^n|} = \lim_{n\to\infty} \frac{2^{nH(P)}}{2^{n\log(|\mathcal{X}|)}} = 0.$$

2. *Despite being 'small', the probability of $\mathcal{T}_\epsilon^{(n)}(P)$ is arbitrarily close to 1 for large n. This means that if we draw an i.i.d. sequence of length n with respect to P, then with arbitrarily high probability this sequence lands in the set $\mathcal{T}_\epsilon^{(n)}(P)$.*

3. *Inside this 'small and high probability' set, all sequences are roughly equiprobable (no typical sequence is favorable over another). Indeed, for all $x^n \in \mathcal{T}_\epsilon^{(n)}(P)$*

$$P^{\otimes n}(\{x^n\}) \doteq 2^{-nH(P)} \doteq \frac{1}{\left|\mathcal{T}_\epsilon^{(n)}(P)\right|},$$

which is approximately the uniform distribution on $\mathcal{T}_\epsilon^{(n)}(P)$.