Recap

Letter Typical Sequences:

- We are looking for a subset $T^{(n)}(P) \subseteq \mathcal{X}^n$ such that

(i) "small": $\lim_{n\to\infty} \frac{|T^{(n)}(P)|}{|\mathcal{X}^n|} = 0$

(ii) "high probability": $\lim_{n\to\infty} P^{\otimes n}\left(X^n \in T^{(n)}(P)\right) = 1$

Key Observation:

While for a given $P \in \mathcal{P}(\mathcal{X})$ it may be the case that $P^{\otimes n}(\{x^n\}) > 0, \forall\, x^n \in \mathcal{X}^n$, we expect that for, e.g., $\mathrm{Ber}(\alpha)$ we we will most likely draw a sequence w/ roughly $\alpha \cdot n$ ones and $(1-\alpha)\cdot n$ zeros.

Empirical Frequency / PMF

For any $x^n \in \mathcal{X}^n$

$$N_{x^n}(a) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}}, \qquad a \in \mathcal{X}$$

$$\nu_{x^n}(a) := \frac{1}{n} N_{x^n}(a)$$

The above $\nu_{x^n}(a)$ is the empirical frequency of $x^n \in \mathcal{X}^n$.

To quantify "roughly" we will introduce a sleekness parameter $\varepsilon > 0$.

(Note we are only discussing $\mathcal{X}$ for which $|\mathcal{X}| < \infty$)

Definition (Letter Typical Set):

Let $\mathcal{X}$ be a finite alphabet, $P \in \mathcal{P}(\mathcal{X})$ have pmf p, $n \in \mathbb{N}$ and $\varepsilon > 0$. The ε-letter typical set of n-lengthed sequences with respect to P with slackness ε is

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

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

Example

$P = \text{Ber}(0.3)$, $n = 10$, $\varepsilon = \frac{1}{2}$

$$T_{\frac{1}{2}}^{(10)}\left(\text{Ber}(0.3)\right) = \left\{ x^{10} \in \{0,1\}^{10} : \begin{matrix} |\nu_{x^n}(0) - 0.7| \leq 0.35 \\ |\nu_{x^n}(1) - 0.3| \leq 0.15 \end{matrix} \right\}$$

Remark (Initial Observation)

① If $p(a) = 0$, then $\nu_{x^n}(a) = 0 \ \ \forall x^n \in T_\varepsilon^{(n)}(P)$, meaning that zero probability letters cannot appear in letter typical sequences.

② Suppose $\varepsilon < 1$. If $p(a) > 0$, then $\nu_{x^n}(a) > 0 \ \ \forall x^n \in T_\varepsilon^{(n)}(P)$ (indeed, if $\nu_{x^n}(a) = 0$ for some x^n, then we must have $p(a) \leq \varepsilon \cdot p(a)$ which is a contradiction)

③ If $P = \text{Unif}(\mathcal{X})$, then for all $x^n \in T_\varepsilon^{(n)}(P)$,

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

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

The strength of letter typical sequences lie in the fact that they lend themselves well for empirical averages, as shown in the following lemma.

Lemma (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 $x^n \in T_\varepsilon^{(n)}(P)$, we have

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

Proof

$\forall\, x^n \in T_\varepsilon^{(n)}(P)$, $a \in \mathcal{X}$ we have $|\nu_{x^n}(a) - p(a)| \leq \varepsilon p(a)$ 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

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

$$= \varepsilon\, \mathbb{E}_P[g]$$

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

Theorem (Properties of Letter Typical Sequences):

Suppose $T_\varepsilon^{(n)}(P)$ 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 $x^n \in T_\varepsilon^{(n)}(P)$,

$$2^{-n(1+\varepsilon)H(P)} \leq P^{\otimes n}(\{x^n\}) \leq 2^{-n(1-\varepsilon)H(P)},$$

(2) We have

$$\lim_{n\to\infty} P^{\otimes n}\left(T_\varepsilon^{(n)}(P)\right) = 1$$

(3) For n sufficiently large,

$$(1-\varepsilon)2^{n(1-\varepsilon)H(P)} \leq \left|T_\varepsilon^{(n)}(P)\right| \leq 2^{n(1+\varepsilon)H(P)}$$

Proof

① Consequence of the lemma. Let $g(x) = -\log(p(x))$.

Then $\mathbb{E}_P[g(X)] = H(P)$ and $\frac{1}{n}\sum_{i=1}^{n} g(x_i) = \frac{1}{n}\log\frac{1}{P^{\otimes n}(\{x^n\})}$.

Thus

$$(1-\varepsilon)H(P) \leq \frac{1}{n}\log_2\frac{1}{P^{\otimes n}(\{x^n\})} \leq (1+\varepsilon)H(P)$$

Exponentiating all terms gives

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

② By definition we have

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

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

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

Now set $K = |\mathcal{X}|$ and $f_a = \mathbb{1}_{\{a\}}$ $\forall\, a\in\mathcal{X}$. 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

$$\lim_{n\to\infty} P^{\otimes n}(T_\varepsilon^{(n)}(P)) = \lim_{n\to\infty} P^{\otimes n}\left(\bigcap_{a\in\mathcal{X}}\{x^n : |\nu_{x^n}(a) - p(a)| \leq \varepsilon p(a)\}\right) = 1$$

③ Using the fact that $T_\varepsilon^{(n)}(P) \subseteq \mathcal{X}^n$, we get

$$1 = P^{\otimes}(\mathcal{X}^n) \geq P^{\otimes n}(T_\varepsilon^{(n)}(P)) = \sum_{x^n \in T_\varepsilon^{(n)}(P)} P^{\otimes n}(\{x^n\})$$

$$\geq \sum_{x^n \in T_\varepsilon^{(n)}(P)} 2^{-n(1+\varepsilon)H(P)}$$

(follows from ①) $$= \left| T_\varepsilon^{(n)}(P) \right| 2^{-n(1+\varepsilon)H(P)}$$

$$\Rightarrow \left| T_\varepsilon^{(n)}(P) \right| \leq 2^{n(1+\varepsilon)H(P)}$$

For n sufficiently large, by ② we have $1-\varepsilon \leq P^{\otimes n}(T_\varepsilon^{(n)}(P))$.

Then

$$1-\varepsilon \leq \sum_{x^n \in T_\varepsilon^{(n)}(P)} P^{\otimes n}(\{x^n\}) \leq \left| T_\varepsilon^{(n)}(P) \right| 2^{-n(1-\varepsilon)H(P)}$$

$$\Rightarrow \left| T_\varepsilon^{(n)}(P) \right| \geq (1-\varepsilon)\, 2^{n(1-\varepsilon)H(P)}$$

∎

Remark (Refined Result)

A finite sample bound on the probability of $T_\varepsilon^{(n)}(P)$ can be derived using a refined argument (based on Chernoff bounds). This gives

$$1 - \delta_\varepsilon(P, n) \leq P^{\otimes n}(T_\varepsilon^{(n)}(P)) \leq 1$$

where $\delta_\varepsilon(P, n) := 2|\mathcal{X}| e^{-2n\varepsilon^2\mu^2}$ and $\mu := \min_{a \in \mathcal{X} : p(a) > 0} p(a)$.

Note that $\lim_{n\to\infty} \delta_\varepsilon(P,n) = 0$, for any fixed $\varepsilon > 0$.

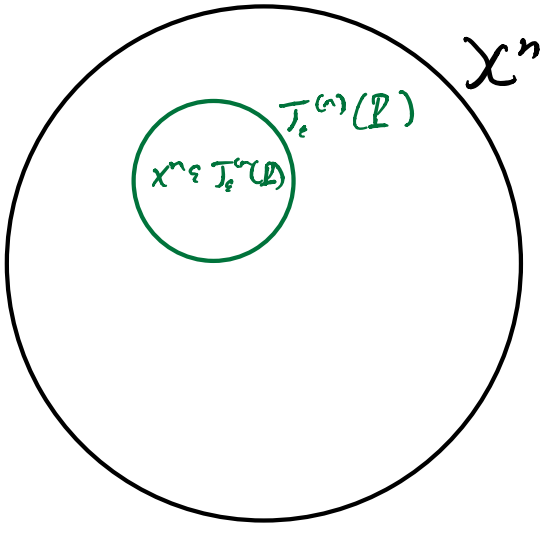
It can thus be shown

$$(1-\delta_\varepsilon(P,n))\, 2^{n(1-\varepsilon)H(P)} \le |T_\varepsilon^{(n)}(P)| \le 2^{n(1+\varepsilon)H(P)}$$

$\forall\, n \in \mathbb{N}$. This strengthens (1), (2) in Theorem above.

Remark (Interpretation & Illustration):

As $n\to\infty$, the set $T_\varepsilon^{(n)}(P)$ 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



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

① The cardinality of $T_\varepsilon^{(n)}(P) \subsetneq \mathcal{X}^n$ is much smaller than that of $\mathcal{X}^n$. More precisely, $\forall\, P \neq \text{Unif}(\mathcal{X})$ and sufficiently small ε, it holds that

$$\lim_{n\to\infty} \frac{|T_\varepsilon^{(n)}(P)|}{|\mathcal{X}^n|} = \lim_{n\to\infty} \frac{2^{nH(P)}}{2^{n\log|\mathcal{X}|}} = 0$$

② Despite being "small", the probability of $T_\varepsilon^{(n)}(P)$ is arbitrarily close to 1 for large n. This means that if we draw an iid sequence of length n with respect to P, then with arbitrarily high probability this sequence lands in the set $T_\varepsilon^{(n)}(P)$.

③ Inside this "small and high probability" set, all sequences are roughly equiprobable (no typical sequence is favorable over another). Indeed, $\forall\, x^n \in T_\varepsilon^{(n)}(P)$

$$P^{\otimes n}(\{x^n\}) \approx 2^{-nH(P)} = \frac{1}{|T_\varepsilon^{(n)}(P)|}$$

which is approximately the uniform distribution on $T_\varepsilon^{(n)}(P)$.