Information Theory for Wireless Communications

Lecture 1:Typical Sequences

Instructor: Dr. Saif K. Mohammed

Scribe: Reza Moosavi

Spring 2012

Consider the following experiment. A computer program is used to generate a binary sequence of length 18 symbols (with probability of zero equal to 2/3). One of the following four sequences is generated from the computer program. Which one is it?

(a) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0, $\Pr(a) = (2/3)^{18} = 6.77 \times 10^{-4}$

(b) 1 0 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 0, $\Pr(b) = (2/3)^9 \cdot (1/3)^9 = 1.32 \times 10^{-6}$

(c) 0 0 0 1 1 0 0 1 0 0 1 1 0 0 0 1 1 0, $\Pr(c) = (2/3)^{11} \cdot (1/3)^7 = 5.28 \times 10^{-6}$

(d) 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1, $\Pr(d) = (1/3)^{18} = 2.58 \times 10^{-9}$.

The answer to this question is sequence (c), even though sequence (a) has a higher occurrence probability. An intuition based reasoning for this phenomena is that, since the source outputs are i.i.d., roughly 2/3 of the 18 symbols should be zero and 1/3 should be one. This is in fact true as we shall see later, and these sequences are called "typical sequences".

I. Stirling's Approximation

Consider again the discrete memoryless binary source X with alphabet $\{0, 1\}$, and let

$$\Pr(X = 0) = p, \qquad \Pr(X = 1) = 1 - p.$$

Let $\{X_n\}$ be a sequence of n independent output of the binary source. Motivated by the above discussion on typical sequences, let us consider the probability of observing a sequence with exactly np zeros (and

hence with $n(1-p)$ ones), which is given by

$$\Pr\left(X_n \text{ has exactly } np \text{ zeros}\right) = \binom{n}{np} p^{np}(1-p)^{n(1-p)} = \frac{n!}{(np)!(n-np)!} p^{np}(1-p)^{n(1-p)},$$

since there are in total $\binom{n}{np}$ independent sequences that have exactly np zeros and the probability of each such sequence is $p^{np}(1-p)^{n(1-p)}$. We would like to find approximately how many such sequences with exactly np zeros exist. This can be done by using the *Stirling's formula*:

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \mathcal{O}(1/n)\right). \tag{1}$$

Using (1) and after appropriate simplifications, we can write

$$\begin{aligned}
\frac{1}{n}\log\binom{n}{np} &= \frac{1}{n}\log\left(\frac{n!}{(np)!(n-np)!}\right) \\
&\approx \frac{1}{n}\log\left(\frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi np}(np/e)^{np}\sqrt{2\pi n(1-p)}\left(n(1-p)\right)^{n(1-p)}}\right) \\
&= -\frac{1}{2n}\log\left(2\pi np(1-p)\right) - p\log p - (1-p)\log(1-p).
\end{aligned}$$

Thus, we have

$$\binom{n}{np} \approx \frac{2^{nH(p)}}{\sqrt{2\pi np(1-p)}}, \tag{2}$$

where $H(p) \triangleq -p\log p - (1-p)\log(1-p)$. It can be shown that $0 \leq H(p) \leq 1$, for any $0 \leq p \leq 1$. Since $H(p) < 1$, then from (2) it appears that the number of typical sequences is a much smaller subset of all possible sequences and further that, with increasing n its size increases as $2^{nH(p)}$. We make this intuition more precise in the next section.

II. Letter Typical Sequences

In this section, we give a formal definition for *letter typicality*, which is sometimes referred to as the strong form of typicality. Let $X \in \mathcal{X} = \{a_1, a_2, \ldots, a_M\}$ be a discrete memoryless source and let $\{X_k\}_{k=1}^n$ be a sequence consisting of n independent source outputs. For notational brevity, we will denote the sequence by X^n. Moreover, let $\Pr\left(X_k = a_i\right) \triangleq p_i$ with $\sum_{i=1}^M p_i = 1$ and let

$$f_i\left(x^n\right) \triangleq \text{number of positions in sequence } x^n \text{ which are equal to } a_i.$$

In other words, $f_i(x^n)$ is the number of a_i:s in the sequence x^n. We now have the following definition for letter typicality:

Definition. For a given $k > 0$, x^n is said to be *k-letter typical* if and only if

$$\left| \frac{f_i(x^n) - np_i}{\sqrt{np_i(1-p_i)}} \right| < k, \quad \forall n, i = 1, 2, \ldots, M. \tag{3}$$

For a given $k > 0$, let us denote the set of all k-letter typical sequences with $\mathcal{T}^k$, that is

$$\mathcal{T}^k \triangleq \{x^n \in \mathcal{X}^n \mid x^n \text{ is } k\text{-letter typical}\}.$$

We have the following theorem.

Theorem 1. *For any given constant $\epsilon > 0$, choose a constant k such that $1/k^2 < \epsilon/M$. Then for any $n \geq 1$, we have:*

1) *The set of k-letter typical sequences of length n has total probability greater than $1 - \epsilon$. In other words, if X^n is the random source output of length n, then*

$$\Pr\left(X^n \in \mathcal{T}^K\right) > 1 - \epsilon.$$

2) *There exists a constant $A > 0$ depending on k and $\{p_i, i = 1, \ldots, M\}$, such that for each k-letter typical sequence x^n,*

$$2^{-nH - A\sqrt{n}} < \Pr(X^n = x^n) < 2^{-nH + A\sqrt{n}},$$

where $H \triangleq -\sum_{i=1}^{M} p_i \log p_i$.

3) *The number of k-letter typical sequences is $2^{n(H+\delta_n)}$, where*

$$\lim_{n\to\infty} \delta_n = 0.$$

Proof:

1) For the first part, we have

$$\begin{aligned}\Pr(X^n \text{ is not a typical sequence}) &= \Pr\left(\left| \frac{f_i(X^n) - np_i}{\sqrt{np_i(1-p_i)}} \right| \geq k, \text{ for at least one } i\right) \\ &\leq \sum_{i=1}^{M} \Pr\left(\left| \frac{f_i(X^n) - np_i}{\sqrt{np_i(1-p_i)}} \right| \geq k\right) \overset{(*)}{\leq} \sum_{i=1}^{M} \frac{1}{k^2} = \frac{M}{k^2} < \epsilon,\end{aligned}$$

where in step $(*)$, we have used the Chebyshev's inequality. More precisely, for random variable Y with expected value μ_Y and standard deviation σ_Y, we have according to the Chebyshev's inequality

$$\Pr\left(|Y-\mu_Y| \geq k\sigma_Y\right) \leq \frac{1}{k^2}.$$

Note that in this case $Y = f_i\left(X^n\right)$, for which it is easy to see that $\mu_Y = np_i$ and that $\sigma_Y = \sqrt{np_i(1-p_i)}$.

2) Suppose that x^n is k-letter typical, then from the definition, we know that

$$np_i - k\sqrt{np_i(1-p_i)} < f_i\left(x^n\right) < np_i + k\sqrt{np_i(1-p_i)}. \tag{4}$$

Since X^n consists of i.i.d. random variables, we have

$$\Pr\left(X^n = x^n\right) = \Pr\left(X_1 = x_1\right)\Pr\left(X_2 = x_2\right)\ldots\Pr\left(X_n = x_n\right) = \prod_{i=1}^{M} p_i^{f_i(x^n)}$$

and thus

$$-\log\Pr\left(X^n = x^n\right) = -\sum_{i=1}^{M} f_i\left(x^n\right)\log p_i.$$

Using the lower and the upper bounds for $f_i\left(x^n\right)$ defined in (4), we get

$$\begin{aligned}-\sum_{i=1}^{M}\left(np_i - k\sqrt{np_i(1-p_i)}\right)\log p_i &< -\log\Pr\left(X^n = x^n\right) \\ &< -\sum_{i=1}^{M}\left(np_i + k\sqrt{np_i(1-p_i)}\right)\log p_i.\end{aligned}$$

Now by defining $A \triangleq -k\sum_{i=1}^{M}\sqrt{p_i(1-p_i)}\log p_i$, we see that

$$nH - A\sqrt{n} < -\log\Pr\left(X^n = x^n\right) < nH + A\sqrt{n}$$

or equivalently,

$$2^{-nH-A\sqrt{n}} < \Pr\left(X^n = x^n\right) < 2^{-nH+A\sqrt{n}}. \tag{5}$$

3) We first note that

$$\sum_{x^n\in\mathcal{T}^k}\Pr\left(X^n = x^n\right) \leq 1. \tag{6}$$

We can also see from (5) that

$$\sum_{x^n \in \mathcal{T}^k} \Pr\left(X^n = x^n\right) \geq |\mathcal{T}^k| 2^{-nH - A\sqrt{n}}.$$

Thus, combining (5) and (6), we have

$$|\mathcal{T}^k| \leq 2^{nH + A\sqrt{n}} = 2^{n\left(H + \frac{A}{\sqrt{n}}\right)} = 2^{n(H + \delta_1(n))}, \tag{7}$$

where $\delta_1(n) \triangleq A/\sqrt{n}$. On the other hand, from the first part of the theorem, we know that

$$\Pr\left(X^n \in \mathcal{T}^k\right) = \sum_{x^n \in \mathcal{T}^k} \Pr\left(X^n = x^n\right) > (1 - \epsilon) \tag{8}$$

and from (5), we have

$$\sum_{x^n \in \mathcal{T}^k} \Pr\left(X^n = x^n\right) \leq |\mathcal{T}^k| 2^{-nH + A\sqrt{n}}. \tag{9}$$

Combining (8) and (9), we get,

$$|\mathcal{T}^k| > (1 - \epsilon) 2^{nH - A\sqrt{n}} = 2^{nH - A\sqrt{n} + \log_2(1-\epsilon)} = 2^{n(H - \delta_2(n))}, \tag{10}$$

with $\delta_2(n) \triangleq \frac{A}{\sqrt{n}} + \frac{\log_2(1-\epsilon)}{n}$. From (7) and (10) it follows that

$$2^{n(H - \delta_2(n))} < |\mathcal{T}^k| < 2^{n(H + \delta_1(n))}.$$

Therefore, $|\mathcal{T}^k|$ must be of the form $|\mathcal{T}^k| = 2^{n(H + \delta_n)}$, where $-\delta_2(n) < \delta_n < \delta_1(n)$. Clearly,

$$\lim_{n \to \infty} \delta_1(n) = \lim_{n \to \infty} \delta_2(n) = 0,$$

and hence

$$\lim_{n \to \infty} \delta_n = 0,$$

and the proof is complete.

■

Theorem 1 essentially means that for a given constant $0 < \epsilon < 1$, the random sequence X^n will belong to $\mathcal{T}^k$ ($k > \sqrt{M/\epsilon}$), with probability greater than $1 - \epsilon$. Therefore, by choosing ϵ small enough, the random source output will belong to the corresponding k-letter typical sequence $\mathcal{T}^k$, with an arbitrarily high probability close to 1. What is more interesting is that for any ϵ, the number of distinct sequences

in $\mathcal{T}^k$ increases exponentially with n as roughly 2^{nH} (which is much smaller than the total number of possible sequences $2^{n\log_2 M}$).

The reason for this interesting phenomenon can be explained as follows. Consider the random variable, $Y_i = \frac{f_i(X^n)}{n}$ which has a mean p_i and standard deviation $\sqrt{\frac{p_i(1-p_i)}{n}}$. Note that the mean value for Y_i is constant and remains the same as the sequence length n increases, whereas its standard deviation scales as $\sqrt{\frac{1}{n}}$. This means that as we increase n, the probability distribution for Y_i becomes more and more centered around the mean value p_i, as illustrated by figures 1 and 2. This in turn means that for large n, most probability mass is distributed around the mean value. Therefore, for large n, the output sequence X^n will have approximately np_i a_i:s. Since the sequences x^n for which $\frac{f_i(x^n)}{n}$ is close to p_i belong to a small subset of all possible sequences in $\mathcal{X}^n$, we conclude that the random output sequence X^n also belongs to a small subset of $\mathcal{X}^n$, which is called the typical set.

III. Entropy Typical Sequences

In this section, we will introduce another "weaker" definition for typicality known as *entropy typicality* and we will give an important result which is known as the asymptotic equipartition property (AEP). As before, let $X \in \mathcal{X} = \{a_1, a_2, \ldots, a_M\}$ be a discrete memoryless source with probability mass function $p(x)$. Moreover, let $p(a_i) = p_i$, with $\sum_{i=1}^{M} p_i = 1$.

Definition. For a given $\epsilon > 0$, a sequence x^n is *ϵ-entropy typical* if

$$\left| \frac{-\log_2 \Pr\left(X^n = x^n\right)}{n} - H \right| < \epsilon, \tag{11}$$

with $H \triangleq -\sum_{i=1}^{M} p_i \log p_i$.

A. Asymptotic Equipartition Property

The asymptotic equipartition property (AEP) is formalized as follows:

Theorem (AEP). *If $X_1, X_2, \ldots, X_n$ are i.i.d. with probability mass function $p(x)$, then*

$$-\frac{1}{n}\log p\left(X_1, \ldots, X_n\right) \xrightarrow[n\to\infty]{\text{in prob.}} H.$$

Proof:

$$-\frac{1}{n}\log p\left(X_1, \ldots, X_n\right) \overset{\text{(i.i.d.)}}{=} -\frac{1}{n}\sum_{i=1}^{n}\log p\left(X_i\right).$$

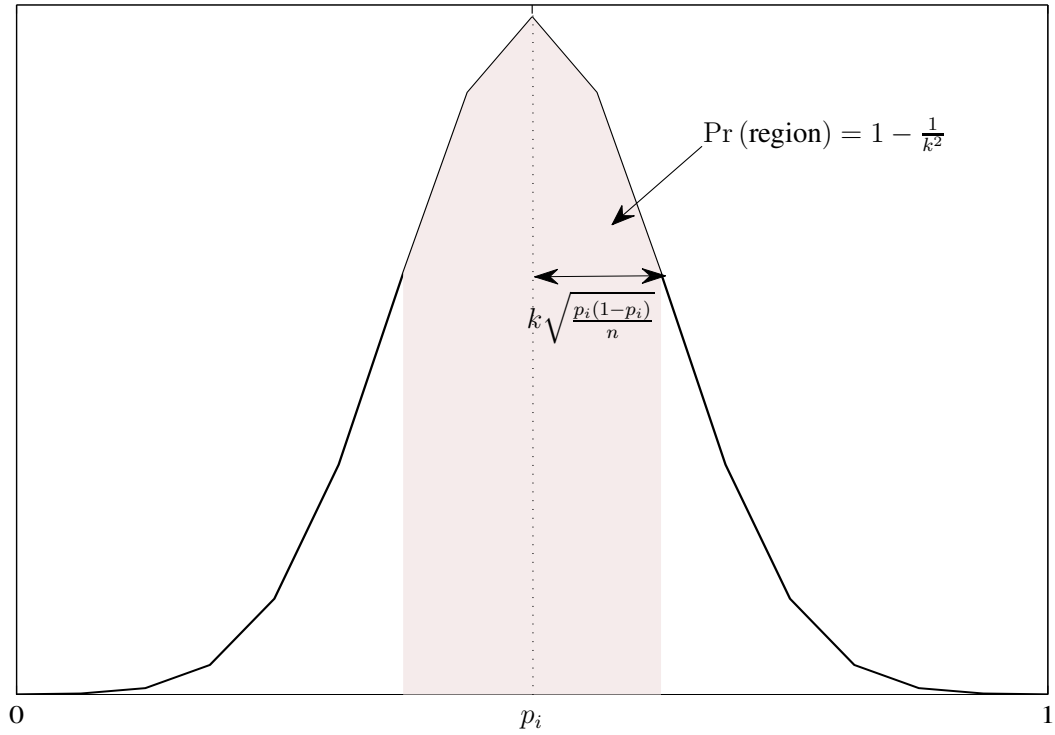

Fig. 1. Normalized mass distribution function for random variable $\frac{f_i(X^n)}{n}$ for $n = 16$, $k = 2$ and $p_i = 1/2$.

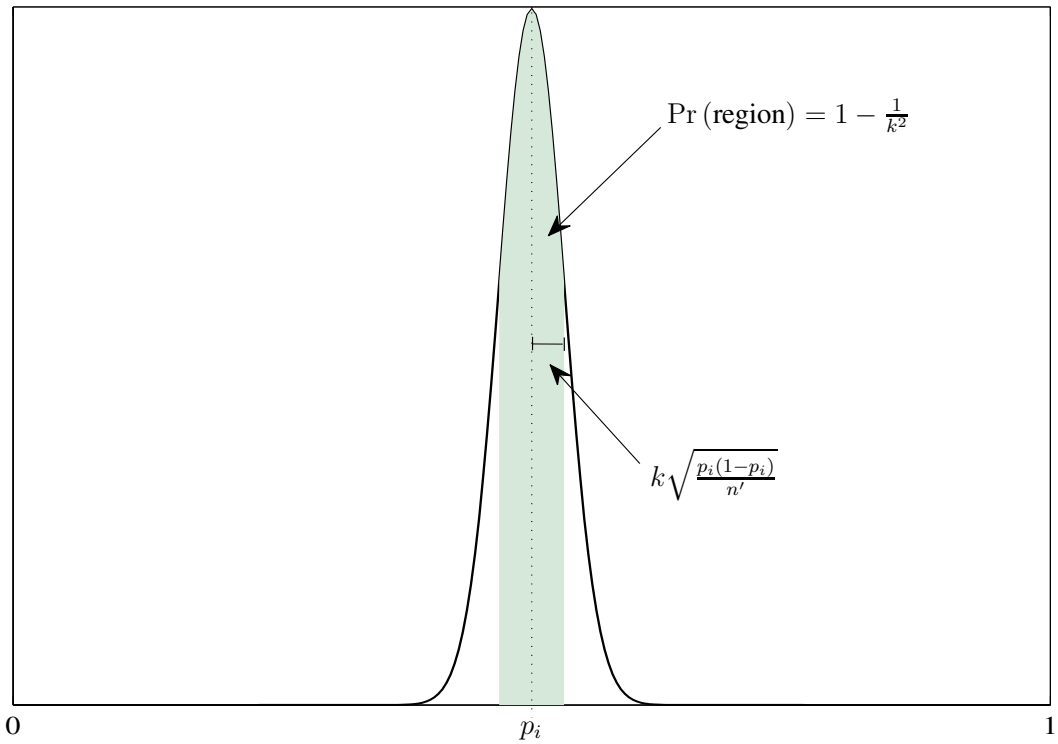

Fig. 2. Normalized mass distribution function for random variable $\frac{f_i\left(X^{n'}\right)}{n'}$ for $n' = 256$, $k = 2$ and $p_i = 1/2$. As we see the distribution is more centered around the mean value, since $n' \gg n$.

Now by defining random variable $\eta_i \triangleq \log p\left(X_i\right)$, and noting that η_i:s are also i.i.d. we conclude from the law of large numbers that as $n \to \infty$,

$$-\frac{1}{n}\sum_{i=1}^{n}\eta_i \overset{\text{in prob.}}{\longrightarrow} -\mathbb{E}\left(\log p(X)\right) = -\sum_{i=1}^{M} p_i \log p_i = H.$$

■

Definition. For a given constant $\epsilon > 0$, the ϵ-typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of all sequences x^n with the property:

$$2^{-n(H+\epsilon)} \leq \Pr\left(X^n = x^n\right) \leq 2^{-n(H-\epsilon)}.$$

In other words,

$$A_\epsilon^{(n)} \triangleq \left\{x^n \mid 2^{-n(H+\epsilon)} \leq \Pr\left(X^n = x^n\right) \leq 2^{-n(H-\epsilon)}\right\}.$$

We next give a theorem which describes the properties of the typical set $A_\epsilon^{(n)}$.

Theorem 2. *If $x^n \in A_\epsilon^{(n)}$, then*

1) $H - \epsilon \leq -\frac{1}{n}\log \Pr\left(X^n = x^n\right) \leq H + \epsilon$,
2) $\Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$, *for sufficiently large* n,
3) $|A_\epsilon^{(n)}| < 2^{n(H+\epsilon)}$, *and*
4) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H-\epsilon)}$, *for sufficiently large* n.

Proof:

1) The first part follows directly from the definition for $A_\epsilon^{(n)}$.
2) From the AEP theorem, we know that

$$-\frac{1}{n}\log p\left(X^n\right) \longrightarrow H, \quad \text{in probability}$$

as $n \to \infty$. Using the definition of the convergence in probability, we conclude that for any $\delta > 0$, there exists a positive integer n_0, such that for all $n > n_0$,

$$\Pr\left(\left|-\frac{\log \Pr\left(X^n\right)}{n} - H\right| < \epsilon\right) > (1-\delta).$$

By setting $\delta = \epsilon$ and noting that all sequences in typical set $A_\epsilon^{(n)}$ satisfy

$$\left|-\frac{\log \Pr\left(X^n\right)}{n} - H\right| < \epsilon,$$

we conclude that $\Pr\left(A_\epsilon^{(n)}\right) > 1-\epsilon$ for sufficiently large n.

3) Using

$$1 = \sum_{x^n} \Pr\left(X^n = x^n\right) \geq \sum_{x^n \in A_\epsilon^{(n)}} \Pr\left(X^n = x^n\right) > \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H+\epsilon)} = |A_\epsilon^{(n)}| 2^{-n(H+\epsilon)},$$

we see that $|A_\epsilon^{(n)}| < 2^{n(H+\epsilon)}$.

4) Using the second part of the theorem, we know that for sufficiently large n,

$$(1-\epsilon) < \Pr\left(A_\epsilon^{(n)}\right) = \sum_{x^n \in A_\epsilon^{(n)}} \Pr\left(X^n = x^n\right) \leq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H-\epsilon)} = |A_\epsilon^{(n)}| 2^{-n(H-\epsilon)},$$

and thus $|A_\epsilon^{(n)}| > (1-\epsilon)2^{n(H-\epsilon)}$.

■

IV. High Probability Set

From Theorem 3, we know that the ϵ-typical set $A_\epsilon^{(n)}$ is a fairly small set that has most of the probability. But is there a smaller set with such property? This section answers this question.

Definition. For a constant $\delta > 0$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest-size set, such that

$$\Pr\left(X^n \in B_\delta^{(n)}\right) \geq 1-\delta.$$

We next prove that the typical set has essentially the same number of elements as the smallest set $B_\delta^{(n)}$, to first order in the exponent.

Theorem 3. *Let $\delta < 1/2$. For any $\delta' > 0$:*

$$\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta',$$

for sufficiently large n.

According to this theorem, for sufficiently large n (depending on δ and δ'), $B_\delta^{(n)}$ has at least $2^{n(H-\delta')}$ elements. We also know that $A_\epsilon^{(n)}$ has about $2^{n(H\pm\epsilon)}$ elements. Therefore, $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ have roughly the same number of elements to first order in the exponent.

Proof: Consider the ϵ-typical set $A_\epsilon^{(n)}$ and the high probability set $B_\delta^{(n)}$. According to Theorem 2,

there exists a positive integer $n_0(\epsilon)$ such that for $n \geq n_0(\epsilon)$,

$$\Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$$

Then we can write,

$$\begin{aligned}\Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) &= \Pr\left(A_\epsilon^{(n)}\right) + \Pr\left(B_\delta^{(n)}\right) - \Pr\left(A_\epsilon^{(n)} \cup B_\delta^{(n)}\right) \\ &> (1-\epsilon) + (1-\delta) - 1 = 1 - \epsilon - \delta, \qquad \text{for } n \geq n_0(\epsilon).\end{aligned}$$

Since we know that

$$\Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) = \sum_{x^n \in \left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right)} \Pr\left(X^n = x^n\right),$$

and using the upper bound for $\Pr\left(X^n = x^n\right)$ that we found in Theorem 2, we have,

$$\begin{aligned}1 - \epsilon - \delta < \Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) &= \sum_{x^n \in \left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right)} \Pr\left(X^n = x^n\right) \\ &\leq \sum_{x^n \in \left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right)} 2^{-n(H-\epsilon)} = \left|A_\epsilon^{(n)} \cap B_\delta^{(n)}\right| 2^{-n(H-\epsilon)} \leq \left|B_\delta^{(n)}\right| 2^{-n(H-\epsilon)}.\end{aligned}$$

Therefore,

$$|B_\delta^{(n)}| > (1 - \epsilon - \delta)\, 2^{n(H-\epsilon)}, \qquad \text{for } n \geq n_0(\epsilon).$$

Finally, we choose $\epsilon \leq \min\{1-\delta, \delta'\}$ and we define

$$n_1 \triangleq \frac{-\log(1-\delta-\epsilon)}{\delta' - \epsilon}.$$

Note that n_1 does not need to be an integer. Clearly, for $n \geq n_1$, we have that,

$$\frac{-\log(1-\delta-\epsilon)}{n} \leq \frac{-\log(1-\delta-\epsilon)}{n_1} = \delta' - \epsilon,$$

and thus

$$\frac{\log(1-\delta-\epsilon)}{n} \geq \epsilon - \delta'.$$

Using the above parameters, we see that for $n \geq \max\{n_0(\epsilon), n_1\}$ we can write

$$\begin{aligned} |B_\delta^{(n)}| &> (1-\epsilon-\delta)\, 2^{n(H-\epsilon)} = 2^{n\left(H-\epsilon+\frac{\log(1-\delta-\epsilon)}{n}\right)} \\ &\geq 2^{n(H-\epsilon+(\epsilon-\delta'))} = 2^{n(H-\delta')}, \end{aligned}$$

or equivalently,

$$\frac{1}{n}\log|B_\delta^{(n)}| > H-\delta'.$$

■