---

commentable: true protected: numbering: type: repopath: mathjax: true categories: Basic-IT tags: Basic-IT keywords: Typical-set RV description: Notes on Typicality mermaid: true highlight: true status: Archive

Typicality

In 0-error data compression, we proved that the infimum of expected length of codewords is equal to the entropy. We go further to explore entropy in therms of the asymptotic behavior of i.i.d. sequences, known as the **Asymptotic Equipartition Property** (AEP).

We consder a diecrete random sequence with each of the variables $X_k, k \geq 1$in a finite alphabet X with probability distribution p_X. All of the Nvariables are i.i.d., i.e.,

$$X^n = (X_1, X_2, \cdots, X_n) \tag{1}$$

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_{X_i}(x_i) \tag{2}$$

$$p_{X_i}(x_i) = p_X(x_i), \forall i = 1, 2, \cdots, n, \tag{3}$$

sad

where we note $p_X : \mathrm{X} \rightarrow [0, 1]$the distribution of X, and (here) indicates that all the possibiility of $X_i, \forall i = [N]$, is the same as $P_{X_i} = P_X$.* Note: In the following relations, we use $P_X$to denote the distribution of any X(rather than X^n), and do not care whose distribution it is.

In the context, the logarithms are in the base 2 by default.

Equality and Convergence of RV

There are several senses that 2RVs can be considered to be equivalent.

Definition 1: (Equivalence of RVs):\ Let X, Y be 2 RVs. We say $X$and Y are:

- equal in distribution (denoted $X \overset{d}{=} Y$) iff they have the same CDF:

$$Pr\{X \leq x\} = Pr\{Y \leq x\}, \quad \forall x \in \mathrm{X} \tag{4}$$

- equal almost surely (denoted $X \overset{a.s.}{=} Y$) iff $Pr\{X = Y\} = 0$.
- equal iff $X(\omega) = Y(\omega), \quad \forall \omega \in \Omega$.

#

- Note: For practical purpose, the measure spaces of $X$and Y are rarely explicitly characterized, so real equality is the least useful while the notion of almost sure equality is as strong as the actual equality.

Lemma 1: Let X, Y be RVs.\ 1. $M_X(t) = M_Y(t), \forall t \in \mathrm{R}$, then $X \overset{d}{=} Y$.\ 2. $X \overset{a.s.}{=} Y \iff d_\infty(X; Y) \overset{\Delta}{=} \text{ess sup}_\omega |X(\omega) - Y(\omega)| = 0$. #

The convergence of RV describes that a sequence of essentially random or unpredictable events (characterized by $\{X_n\}_{n \in \mathrm{N}_+}$) can be settled into patterns that are intuitively

unchanging.

Definition 2: (Convergence in Distribution):\ A sequence $X_1, X_2, \cdots$of RVs is said to converge in distribution, or converge weakly to a RV X, denoted $X_n \xrightarrow{d} X$, or $X_n \Rightarrow X$iff $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$holds for all $x \in \mathrm{X}$ at which $F_X$is continuous. # The constraints on continuous points are essential. Here is a counterexample:

Example 1: Consider $\{X_n\}_{n\in\mathrm{N}_+}$that $X_n \sim U([1, 1/n])$then the CDF converges to a degenerate RV $X = 0$. The contradictory is that $F_X(0) = 1$while $\forall n \in \mathrm{N}_+, \quad F_{X_n}(0) = 0,$ not convergence at the discontinuous point $x = 0$. #

weak convergence do not talk about the independence or correlated relations. Here are some lemmas related:

Lemma 2: (Scheff's lemma): Let f_na sequence of integrable function on a measure space (X, Σ, μ)that converges a.e. to another integrable function f, then:

$$\int |f_n - f| d\mu \to 0 \iff \int |f_n| d\mu \to \int |f| d\mu \tag{5}$$

So the a.e. pointwise convergence in PDF implies convergence in distribution. #

- Note: The counterpart is not true. Consider $f_{X_n}(x) = (1 - \cos(2n\pi x)) I_{[0,1]}(x)$where I_A is the indicator function of A. Then $X_n \xrightarrow{d} X \sim U([0, 1])$while the PDF shows no convergence.

Lemma 3: (Levy's continuity theorem):\ The sequence $\{X_n\}$converges in distribution to Xiff the sequence of corresponding characteristic functions $\{\phi_n\}$converges pointwise to the characteristic function $\phi$of X. #

Definition 3: (Convergence in Probability):\ A sequence $X_1, X_2, \cdots$of RVs is said to converge in probability to a RV X, denoted $X_n \xrightarrow{p} X$iff

$$\forall \delta > 0, \quad \lim_{n\to\infty} Pr\{|X_n - X| > \delta\} = 0 \tag{6}$$

#

- Note: for metric space, it is $\forall \delta > 0, \quad \lim_{n\to\infty} Pr\{d(X_n - X) > \delta\} = 0.\}$

Lemma 4: The following 3 interpreters of convergence in probability is equivalent:

- $X_n \xrightarrow{p} X$
- $\forall \epsilon > 0, \exists N \in \mathrm{N}_+, \forall n \geq N, \forall \delta > 0, Pr\{|X_n - X| > \delta\} < \epsilon.$

- $\forall \epsilon > 0, \exists N \in \mathrm{N}_+, \forall n \geq N, Pr\{|X_n - X| > \epsilon\} < \epsilon.$

#

Proof: $1 \iff 2$by the definition of limitation, $2 \rightarrow 3$by letting $\delta = \epsilon$. We prove $3 \rightarrow 2$. To prove 2 holds $\forall \delta > 0$, we have the following 3 cases:

- $\delta = \epsilon$. From 3 we get 2 immediately.
- $\delta < \epsilon$. As $\delta > 0$, by 3 $\exists N \in \mathrm{N}_+, \forall n \geq N, Pr\{|X_n - X| > \delta\} < \delta < \epsilon.$
- $\delta > \epsilon$. Then $\exists N \in \mathrm{N}_+, \forall n \geq N, Pr\{|X_n - X| > \delta\} < Pr\{|X_n - X| > \epsilon\} < \epsilon.$

So $2 \iff 3$. #

Definition 4: (Convergence Almost Everywhere):\ A sequence $X_1, X_2, \cdots$of RVs is said to converge almost everywhere, or converges with probability 1 to a RV X, denoted $X_n \xrightarrow{a.s.} X$ iff

$$Pr\{\lim_{n\rightarrow\infty} X_n = X\} = 1 \tag{7}$$

or equivalently,

$$P\left(\{\omega \in \Omega : \lim_{n\rightarrow\infty} X_n(\omega) = X(\omega)\}\right) = 1 \tag{8}$$

$$\forall \epsilon > 0, P\left(\limsup_{n\rightarrow\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0 \tag{9}$$

#

Definition 5: (Convergence Everywhere):\ A sequence $X_1, X_2, \cdots$of RVs is said to converge everywhere, or converges pointwosely to a RV X, denoted $X_n \rightarrow X$iff

$$\lim_{n\rightarrow\infty} X_n = X \tag{10}$$

or equivalently,

$$\{\omega \in \Omega : \lim_{n\rightarrow\infty} X_n(\omega) = X(\omega)\} = \Omega \tag{11}$$

#

Definition 6: (Convergence in Mean):\ Let $r \geq 1$, a sequence $X_1, X_2, \cdots$of RVs is said to converge in the r-th mean, or in the L^r-norm to a RV X, denoted $X_n \xrightarrow{L^r} X$iff the r-th absolute moments $E[|X_n|^r]$, and $E[|X|^r]$exists, and

$$\lim_{n\rightarrow\infty} E[|X_n - X|^r] = 0 \tag{12}$$

#

Lemma 5: (implimentations of RV convergences):\ The relations between the above 4 kinds of convergences are shown as follows:

- Convergence in probability implies convergence in distribution, but convergence in distribution implies convergence in probability when the limiting random variable X is a constant.
- Almost sure convergence implies convergence in probability, but convergence in probability does not imply almost sure convergence.
- By Markov's inequality $Pr\{X \geq a\} \leq \frac{E[X]}{a}$, convergence in the r-th mean, for $r \geq 1$, implies convergence in probability.
- if $r > s \geq 1$, convergence in r-th mean implies convergence in s-th mean.
- Sure convergence of a random variable implies all the other kinds of convergence stated above.

#

Law of Large Numbers

We review the definition of LLN, performance of large times of experiments that the average tends to the expected value (when it exists).* Note: The average of trials may not converge in some cases because of heavy tails, such as Cauchy distribution (where it has no expectation) and Pareto distribution when $\alpha < 1$(where it has infinite expectation). We mainly talk about i.i.d. Lebesgue integrable RV sequence $\{X_k\}_{k\in\mathrm{N}}$generated by X, where $E[X] = \mu$, $Var(X) = \sigma^2$

- Note: Lebesgue integrability of $X_k$means that the expected value $E[X_k]$exists according to Lebesgue integration and is finite. It does not mean that the associated probability measure is absolutely continuous w.r.t. Lebesgue measure.

Theorem 1: (Khinchin's weak LLN): The sample average converges in probability towards the expected value

$$\bar{X}_n \xrightarrow{p} \mu, \quad n \to \infty \tag{13}$$

#

Actually, finite variance is not necessary though it may simplify the related proof. Large or infinite variance may make the convergence slower but LLN also holds. The convergence of $\bar{X}$to a degenerate RV requires $Var(\bar{X}) = \sigma^2/n$to converge to 0, which may be true even

if $\sigma^2 \to \infty$. In Chebyshev's weak LLN, we release the restriction of i.i.d. sequence and finite variance:

Theorem 2: (Chebyshev's weak LLN): For independent Lebesgue integrable RV sequence $\{X_k\}_{k\in\mathrm{N}}$with equal expected value μ, if $\lim_{n\to\infty} Var(\bar{X}_n) = 0$, then:

$$\bar{X}_n \xrightarrow{p} \mu, \quad n \to \infty \tag{14}$$

Reminds: weak LLN may hold when the expectation do not exist.

Example 2: Let $\{X_k\}_{k\in\mathrm{N}}$be an independent zero-mean Gaussian RV sequence with $Var(X_k) = \frac{2n}{\log(n+1)}$, which is not bounded.

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{\log n} \to 0 \tag{15}$$

which satisfies the weak LLN. #

Theorem 3: (Kolmogorov's strong LLN):

- $\bar{X}_n \xrightarrow{a.s.} \mu, \quad n \to \infty$
- when $\{X_k\}_{k\in\mathrm{N}}$ are not i.i.d, then $\bar{X}_n - E[\bar{X}_n] \xrightarrow{a.s.}$, provided that $\sum_{k=1}^{\infty} \frac{1}{k^2} Var(X_k) < \infty$ and $\forall k \in \mathrm{N}, X_k$has finite second moment.

#

- Note: The strong Law is seen to be a special case of the pointwise ergodic theorem
- Note: The strong LLN can derive the weak one but do not hold in cases when the expected value do not exist.

The differences between the weak law and the strong law are in that:

- The weak law states that for a specified large n , the average $\bar{X}_n$is likely to be near μ. Thus, it leaves open the possibility that $\left|\bar{X}_n - \mu\right| > \epsilon$happens an infinite number of times, although at infrequent intervals. (Not necessarily $\left|\bar{X}_n - \mu\right| = 0$for all n).
- The strong law shows that this almost surely will not occur. In particular, it implies that with probability 1, we have that for any $\epsilon > 0$the inequality $\left|\bar{X}_n - \mu\right| < \epsilon$holds for all large enough n.

The strong law does not hold in the following cases, but the weak law does.

Example 3:

- Let Xbe an exponentially distributed random variable with parameter 1 . The random variable $\sin Xe^X X^{-1}$has no expected value according to Lebesgue integration, but using conditional convergence and interpreting the integral as a Dirichlet integral, which is an improper Riemann integral, we can say:

$$E\left(\frac{\sin(X)e^X}{X}\right) = \int_0^\infty \frac{\sin(x)e^x}{x} e^{-x} dx = \frac{\pi}{2} \tag{16}$$

- Let $x$be geometric distribution with probability $p = 1/2$. The random variable $2^X(-1)^X X^{-1}$does not have an expected value in the conventional sense because the infinite series is not absolutely convergent, but using conditional convergence, we can say:

$$E\left(\frac{2^X(-1)^X}{X}\right) = \sum_1^\infty \frac{2^x(-1)^x}{x} 2^{-x} = -\ln(2) \tag{17}$$

- If the cumulative distribution function of a random variable is

$$\begin{aligned} 1 - F(x) &= \tfrac{e}{2x\ln(x)}, x \geq e \\ F(x) &= \tfrac{e}{-2x\ln(-x)}, x \leq -e \end{aligned} \tag{18}$$

then it has no expected value, but the weak law is true.

For sequence of i.i.d. functions of RV: $\{f(X_k, \theta)\}_{k\in\mathrm{N}}$with parameter $\theta \in \Theta$, by the strong LLN we know that for fixed θ, the sample mean converges to $E[f(X, \theta)]$. This is known as the pointwise convergence in θ. To make the convergence happens uniformly in θ, we have the **Uniform** LLN.

Theorem 4: (Uniform LNN): For i.i.d. functions of RV: $\{f(X_k, \theta)\}_{k\in\mathrm{N}}$with parameter $\theta \in \Theta$ where $E[f(X, \theta)]$exists for all θ, if :

- $\Theta$is compact
- $f(x, \theta)$is continuous at each $\theta \in \Theta$for almost all $x \in \mathrm{X}$, and measurable function of x at each θ
- there exists a dominating function $d(x)$such that $E[d(X)] < \infty$, and

$$\|f(x, \theta)\| \leq d(x) \quad \text{for all } \theta \in \Theta \tag{19}$$

Then $E[f(X, \theta)]$is continuous in θ, and

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) - \mathrm{E}[f(X, \theta)] \right\| \xrightarrow{\text{a.s.}} 0 \tag{20}$$

This result is useful to derive consistency of a large class of estimators known as Extremum estimator.

Moreover, an intuitive notion of probability as a long-run relative frequency leads us to the Borel LNN:

Theorem 5:(Borel LLN): If an experiment $(\Omega, \mathrm{B}, \mathbf{P})$is repeated a large number of times independently and under identical conditions, then

$$n \to \infty, \frac{N_n(A)}{n} \xrightarrow{a.s.} P(A), \quad \forall A \in \mathrm{B} \tag{21}$$

#

Lemma 6: (Chebyshev's Inequality):

$$Pr\{|X - E[X] \geq k\sigma\} \leq \frac{1}{k^2}, \quad \forall k > 0, \sigma^2 = Var(X). \tag{22}$$

#

Lemma 7: (Markov's Inequality): Let Xbe a non-negative RV, then

$$Pr\{X \geq a\} \leq \frac{E[X]}{a} \tag{23}$$

#

An intuition to Markov's Inequality is that $E[X] = Pr\{X < a\}E[X|X < a] + Pr\{X \geq a\}E[X|X \geq a]$. So when $E[X|X < a] \geq 0$where $X > 0$, for any possible distribution p_X, the value $E[X|X \geq a] \geq a$. Thus $E[X] \geq Pr\{X \geq a\}E[X|X \geq a] \geq Pr\{X \geq a\}a$.

Weak AEP

Proof of Weak AEP

Theorem 6: (Weak AEP 1):

$$n \to \infty, \quad -\frac{1}{n}p(X^n) \xrightarrow{p} H(X) \tag{24}$$

or equivalently, let $\mathrm{W}_{[X]\epsilon}^n \triangleq \{x^n \in \mathrm{X}^n : |-\frac{\log p(x^n)}{n} - H(X)| \leq \epsilon\}$, then

- $\forall x^n \in \mathrm{W}_{[X]\epsilon}^n, \quad 2^{-n(H(X)+\epsilon)} \leq p_(x) \leq 2^{-n(H(X)-\epsilon)}$;

- $\forall \epsilon > 0, \exists N \in \mathrm{N}_+, \forall n > N, Pr\{X^n \in \mathrm{W}^n_{[X]\epsilon}\} > 1 - \epsilon.$
- $\forall \epsilon > 0, \exists N \in \mathrm{N}_+, \forall n > N, (1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathrm{W}^n_{[X]\epsilon}| \leq 2^{n(H(X)+\epsilon)}.$

We call the set $\mathrm{W}^n_{[X]\epsilon}$the weekly typical set, $x^n \in \mathrm{W}^n_{[X]\epsilon}$the weakly typical sequence, and $-\frac{\log p(x^n)}{n}$the empirical entropy of sequence x^n. #

Proof: By the week law of large numbers of i.i.d. sources, $n \to \infty, -\frac{1}{n}\log p(X^n) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i) \xrightarrow{p} -E[\log p(X)] = H(X)$. #

The weak AEP shows that a small set (compared to X^n, $\mathrm{W}^n_{[X]\epsilon}$is exponentially small) contains almost all probability of $\mathrm{X}^n$with nearly uniform distribution. So randomly choose a sequence from X^n, the probability are near $\frac{1}{|\mathrm{W}^n_{[X]\epsilon}|}$with high probability.* Note: The weekly typical sequence is different from the most likely sequence. It is not necessary fora a weakly typical set to include most likely sequences to possess the probability 1.

Extension of weak AEP: Shannon-McMillan-Breiman(SMB) Theorem

Theorem 7: (Cramer theorem):\ Let $\{X_k\}_{k\in\mathrm{N}}$be a sequence of i.i.d. RVs with finite logarithmic moment generating functuion $\Lambda(t) = \log E[e^{tX}] < \infty$, then the Legendre transform of Λsatisfies:

$$\lim_{n\to\infty} \frac{-\log(Pr\{\sum_{k=1}^n X_k \geq nx\})}{n} = \Lambda^*(x) \triangleq \sup_{t\in\mathrm{R}}(tx - \Lambda(t)) \tag{25}$$

#

- Note: This intuitively states that the probability of a large deviation from mean decays exponentially with number of samples n.

Theorem 8: (SMB Theorem):\ Let $\{X_n\}$be a stationary ergodic process defined on a probability space $(\Omega, \mathrm{B}, \mathbf{P})$, then the weak AEP for $\{X_n\}$shows that

$$n \to \infty, \quad -\frac{1}{n}\log p(X^n) \xrightarrow{a.e.} H \tag{26}$$

where $H$is the entropy rate. #

The assumptions of stationarity/ergodicity/identical of RV is not necessary for the AEP to hold. An intuitive idea is that a form that LLN holds may be applied to weak AEP.

Theorem 9: For independent source $\{X_k\}_{k\in\mathrm{N}}$with bounded $Var(\log p(X_i))$, then the weak AEP holds:

$$n \to \infty, \quad -\frac{1}{n}\log p(X^n) \xrightarrow{p} \bar{H}(X) \triangleq \frac{1}{n}H(X_1, \cdots, X_n) \tag{27}$$

#

Strong AEP and method of type

We introduce a new kind of AEP, which is related to Borel's LLN. The following relations are **called the method of types**, as the term **type** is given as an equivalence relation and the probability of the equivalence class gives the definition of typical set and the corresponding strong AEP:

Cross Entropy and type

Definition 7: (Empirical Distribution):\ For a given random sequence $x^n \in \mathrm{X}^n$, The number of appearances of a in the sequence:

$$N(a|x^n) = \sum_{k=1}^{n} \delta_{x_i,a}, \quad \forall a \in \mathrm{X} \tag{28}$$

where $\delta_{a,b}$is the Kronecker symbol and returns 1ony if $a = b$. We denote:

$$Q(a) = P_{x^n}(a) = \frac{1}{n}N(a|x^n), \tag{29}$$

2

as the empirical distribution of any given x^n, and from (here), empirical distribution Qcan also be seen as the frequency P_{x^n}. #

For i.i.d. sequence $X^n$with distribution p_X, and empirical distribution P_{x^n}in a specific $x^n \in \mathrm{X}^n$, we calculate the $Pr(X^n = x^n)$.

$$\frac{1}{n} log Pr(X^n = x^n) \tag{30}$$

$$= \frac{1}{n} log \prod_{i=1}^{n} P_X(x_i) \tag{31}$$

$$= \frac{1}{n} log \prod_{a \in \mathrm{X}} P_X(a)^{N(a|x^n)} \tag{32}$$

$$= \sum_{a \in \mathrm{X}} \frac{N(a|x^n)}{n} log P_X(a) \tag{33}$$

$$= \sum_{a \in \mathrm{X}} Q(a) log P_X(a) < aid = "CrossEntropy" > 3 < /a > \tag{34}$$

This indicates that the probability of a random sequence only depends on the distribution of each random variable $X$and the frequency of the value x^n. Like the definition of emprical entropy in the weak AEP, we define the cross entropy:

Definition 8: (Cross Entropy):\ We define the cross entropy in an i.i.d. trial with n implements of RV Xas:

$$H(Q, P_X) = - \sum_{a \in \mathrm{X}} Q(a) log P_X(a), < aid = "DefCrossEntropy" > 4 < /a > \tag{35}$$

#

From the definition ([here](#Def Cross Entropy)),

$$Pr(X^n = x^n) = q^{-nH(Q,P_X)} \tag{36}$$

Easy to see that if all the frequency of $a \in \mathrm{X}$on different x^ns are the same, the values of probability distribution of X^non point x^ns are the same. This allows us to divide the sample space X^ninto different parts and each part has the same empirical distribution, and in each part, sequences are automatically uniformly distributed. The partition is done only by Q : $\mathrm{X} \rightarrow \frac{[n]}{n}$. For a formal definition,

Definition 9: (type):\ For an i.i.d. trial of $n$complements of RV Xover a finite alphabet X, the type w.r.t. emperical distribution $Q$is defined by:

$$\mathrm{T}_Q^n = \{x^n \in X^n | P_{x^n} = Q\} \tag{37}$$

so any type $\mathrm{T}_Q^n \subseteq \mathrm{X}^n$is a set of sequences whose probability are numberially the same. # By combinatorial theory, the number of types is the munber of possible empirical distributions:

$$\binom{n+|\mathrm{X}|-1}{n} \leq (n+1)^{|\mathrm{X}|} \tag{38}$$

which is a ploynomial of n* Note: The question of the number of types is equivalent to the question, in which n different balls are independently labeled with $|\mathrm{X}|$labels, or we choose n unordered balls from an urn containing $|\mathrm{X}|$balls with replacement.}

Definition 10: (Rate of a Quantity):\ For any quantity Mwhich is exponentially large or exponentially small, we call the $\frac{1}{n}\log M$the rate of M, where the $\log$takes base of a convenient integer. #

Size of a Type

Lemma 8: (Size of a Type):\ For ni.i.d. experiments over the implement of $X \in \mathrm{X}$, the size of any type satisfies:

$$\frac{1}{n} log|\mathrm{T}_Q^n| \sim H(Q,Q) = H(Q), \quad n \to \infty \tag{39}$$

As the size of type only depends on the empirical distribution $Q = P_{x^n}$, this theorem introduces a natural constant for any given distribution Q:

$$H(Q) = H(Q,Q) = -\sum_{a \in \mathrm{X}} Q(a) log Q(a) \tag{40}$$

which can be an alternative definition of entropy. #

Proof: The size of any type can be shown combinatorially:

$$|\mathrm{T}_Q^n| = \frac{n!}{\prod_{a \in \mathrm{X}} [nQ(a)]!} \tag{41}$$

and noticing the Stirling approximation:

$$n^n \sim n! e^n \to \frac{1}{n} ln n! \sim ln n - 1, \quad n \to \infty \tag{42}$$

We have:

$$\frac{1}{n}log|\mathrm{T}_Q^n| \tag{43}$$

$$\sim(lnn-1)-\sum_{a\in\mathrm{X}}Q(a)(ln[nQ(a)]-1) \tag{44}$$

$$=(lnn-1)-(lnn-1)\sum_{a\in\mathrm{X}}Q(a)-\sum_{a\in\mathrm{X}}Q(a)logQ(a) \tag{45}$$

$$=-\sum_{a\in\mathrm{X}}Q(a)logQ(a)=H(Q,Q),\quad n\to\infty<aid="entr">5 \tag{46}$$

#

It is worth noticing that $\frac{1}{n}log|\mathrm{T}_Q^n|\sim H(Q)$is a exponential approximation, which allows an deviation of a ploynomial times of the size.* Note: This is true that for any ploynomial $P(x)\in\mathrm{R}[x],\lim_{n\to\infty}\frac{e^x}{P(x)}=0\}$. Specifically, for $\mathrm{T}_Q^n$with Qhappens to be the distribution of X, i.e., $Q=P_X$,

$$\frac{1}{\binom{n+|\mathrm{X}|-1}{n}}\le\sum_{x^n\in\mathrm{T}_Q^n}\prod_{i=1}^{n}p_X(x_i)=Q^n(\mathrm{T}_Q^n)\le 1 \tag{47}$$

and by $[here](DefCrossEntropy)$, $\forall x^n\in\mathrm{T}_Q^n,P_{X^n}(x^n)=Q^n(x^n)=q^{-nH(Q)}$. So, we get the two-side bounds of the size of types: * Note: In fact we can choose any possible $P_X$as the number of types are independent with p_X. We choose $P_X=Q$to get a proper lower bound, as the probability will be maximum when $p_X=Q$

$$\frac{q^{nH(Q)}}{\binom{n+|\mathrm{X}|-1}{n}}\le|\mathrm{T}_Q^n|\le q^{nH(Q)},<aid="specsize">6 \tag{48}$$

Probability of a type

Lemma 9: (Probability of a type):\ For ni.i.d. implements of RV $X\sim p$, when $n\to\infty$,

$$\frac{1}{n}\log p_{X^n}(\mathrm{T}_Q^n)\sim D(Q||p) \tag{49}$$

As the probability of type depends on the empirical distribution $Q=P_{x^n}$and real distribution p, this theorem introduces a natural constant for any given distribution $Q$and p:

$$H(Q,p)-H(Q)=D(Q||p)=-\sum_{a\in\mathrm{X}}p(a)log\frac{p(a)}{Q(a)}=D(Q||p)\ge 0 \tag{50}$$

which can be an alternative definition of Kullback-Liebler Divergence (or relative entropy). #

Proof: We can simply calculate the probability of any given type $\mathrm{T}_Q^n$by the multiplication of the probability of any point in the type(as thay are uniformly distributed) and the size of the type, which is given in ([here](#Def Cross Entropy)) and (here):

$$\frac{1}{n}\log P_{X^n}(\mathrm{T}_Q^n) \tag{51}$$
$$=\frac{1}{n}\log[|\mathrm{T}_Q^n|P_{X^n}(x^n : x^n \in \mathrm{T}_Q^n)] \tag{52}$$
$$\sim\frac{1}{n}\log|\mathrm{T}_Q^n| + \frac{1}{n}logP_{X^n}(x^n : P_{x^n} = Q) \tag{53}$$
$$=H(Q) - H(Q,p), \tag{54}$$

7

To put it more specifically from (here),

$$\frac{q^{-nD(Q||p)}}{\binom{n+|\mathrm{X}|-1}{n}} \leq P_{X^n}(\mathrm{T}_Q^n) = |\mathrm{T}_Q^n|P_{X^n}(x^n : x^n \in \mathrm{T}_Q^n) \leq q^{-nD(Q||p)} \tag{55}$$

#

(here) shows that the difference between $H(Q)$and $H(Q,P_X)$(i.e., the difference of empirical distribution and real distribution) determines the probability of a certain type. (Compared to the size of a type, which is only determined by the empirical distribution). We can intuitively cosnsider $D(Q||P_X)$as the "distance" between 2 distributions. The large the distance of $Q$and $P_X$is, the less $P_{X^n}(\mathrm{T}_Q^n)$will be. As a specific case when $P = Q$, $D(Q||P_X) = 0$and the probability gets maximum.

Strong AEP

Although it seems that when $P = Q$, $\frac{1}{n}logP_{X^n}(\mathrm{T}_Q^n) \sim 0$, however, $P_{X^n}(\mathrm{T}_Q^n) \sim 1$ is not correct as the former approximation is in a exponential sense. In fact, when $n \longrightarrow \infty$, the number of types "near" P_X, say, $\exists a \in \mathrm{X}, |Q(a) - P_X(a)| \leq \delta$, is no less than $[2n\delta]$ with nearly (the difference will be bounded in the following statement) the same probability, so the probability of type $P_{X^n}(\mathrm{T}_{P_X}^n) \sim 1/2n\delta \longrightarrow 0, \quad n \longrightarrow \infty$. So we do not care any possibility of any *certain type* but a *range of types*. In this section we show that the possibility concentrates exactly around the maximum type $\mathrm{T}_{P_X}^n$.

- Note: If only $a \in \mathrm{X}$ changes, $Q(a) = \frac{N(a|x^n)}{n} \longrightarrow N(a|x^n) = nQ(a) \in [n(P_X(a) - \delta), n(P_X(a) + \delta)]$, so basically there are no less than $[2n\delta]$different $N(a|x^n)$, and thus $[2n\delta]$different types. }

Definition 11: (Strongly Typical Set): Let $X \sim p$be a RV defined on a finite alphabet X. We define the the union of all "near maximum" types as the *Strongly Typical Set*:

$$\mathrm{T}^n_{[X]_\delta} = \{x^n \in \mathrm{X}^n : ||Q(a) - p_X(a)| \leq \delta, \forall a \in \mathrm{S}_X \wedge Q(a) = 0, \forall a \notin \mathrm{S}_X\} \quad (56)$$

$$= \bigcup_{Q:|Q(a)-P_X(a)|\leq\delta,\forall a\in\mathrm{X},\wedge Q(a)=0,\forall a\notin\mathrm{S}_X} \mathrm{T}^n_Q \quad (57)$$

or alternatively,

$$T^n_{[X]_{\delta'}} = \{x^n \in \mathrm{X}^n : \forall a \notin \mathrm{S}_X, Q(a) = 0 \wedge \sum_{x\in\mathrm{X}} ||Q(x) - p(x)| \leq \delta'\} \quad (58)$$

by letting $\delta' = \sqrt{|\mathrm{X}|}\delta$. The completement of strongly typical set is called the strongly atypical set, denoted by $\mathrm{T}^{nC}_{[X]_\delta} = \mathrm{X}^n - \mathrm{T}^n_{[X]_\delta}$($T^{nC}_{[X]_\delta} = \mathrm{X}^n - T^n_{[X]_\delta}$ resp.). For $x^n \in \mathrm{X}^n$, if $\exists x_i \in \mathrm{X}, |Q(x_i) - p(x_i)| \geq \delta$, then $x^n \in \mathrm{T}^{nC}_{[X]_\delta}$. #

The strong AEP are acknowledged with 3 lemmas, analogous to the weak AEP:

Theorem 10: (Strong AEP I: Size of Strongly Typical Set):

$$\frac{1}{n}|\mathrm{T}^n_{[X]_\delta}| \sim H(p), \quad n \to \infty \quad (59)$$

#

- Note: From [here](#entr we know that the size of type has a $H(p)$exponential approximation. However, The gap between them tends to be infinite, as the probability are tend to 1 and 0 from the following relation.)

Proof: First we give a lemma according to the continuity of $H(\cdot)$:

Lemma 10: due to the uniform continuity of entropy function

$$H : \{(q_1, \cdots, q_{|\mathrm{X}|})|\forall i \in [n], q_i \in [0, 1], \sum_{i=1}^{|\mathrm{X}|} q_i = 1\} \to \mathrm{R}_+, H(Q) = -\sum_{a\in\mathrm{X}} Q(a)logQ(a) \quad (60)$$

we can find $\delta_\epsilon$to make $|H(Q) - H(p)| \leq \epsilon$, i,.e.,

$$\forall \epsilon > 0, \exists \delta_\epsilon > 0, |(q_1, \cdots, q_{|\mathrm{X}|}) - (p_1, \cdots, p_{|\mathrm{X}|})| \leq \delta_1 \to |H(Q) - H(p)| \leq \epsilon \quad (61)$$

#

We pick* Note: Attention that we pick $\delta$from ϵbut not $\epsilon$from δ. This δhas a linear form of ϵ, so they have the same convergence when $n \to \infty$.

$$\delta_\epsilon = min\{\frac{\epsilon}{-\sum_{a\in \mathrm{X}} logp(a)}, \frac{\epsilon}{-\sum_{a\in \mathrm{X}} logQ(a)}\} \tag{62}$$

and $H(Q)$is bounded by:

$$H(Q) \leq H(Q, P_X) = -\sum_{a\in \mathrm{X}} Q(a) log P_X(a) \tag{63}$$

$$\leq -\sum_{a\in \mathrm{X}} (P_X(a) + \delta_\epsilon) log P_X(a) \tag{64}$$

$$\leq -\delta_\epsilon \sum_{a\in \mathrm{X}} log P_X(a) - \sum_{a\in \mathrm{X}} P_X(a) log P_X(a) \tag{65}$$

$$= \epsilon + H(P_X) \tag{66}$$

$$H(Q) = -\sum_{a\in \mathrm{X}} Q(a) log Q(a) \tag{67}$$

$$\geq -\sum_{a\in \mathrm{X}} (P_X(a) - \delta_\epsilon) log Q(a) \tag{68}$$

$$\geq H(P_X) - \epsilon \frac{\sum_{a\in \mathrm{X}} log Q(a)}{\sum_{a\in \mathrm{X}} log Q(a)} \tag{69}$$

$$= H(P_X) - \epsilon \tag{70}$$

$$|Q(a) - P_X(a)| \leq \delta_\epsilon, \forall a \in \mathrm{X} \rightarrow |H(Q) - H(P_X)| \leq \epsilon \tag{71}$$

Which ends the proof of lemma. We then calculate the upper bound of the size of typical set:

$$\frac{1}{n} log |\mathrm{T}^n_{[Q]_{\delta_\epsilon}}| \tag{72}$$

$$= \frac{1}{n} log \sum_{Q:|Q(a)-P_X(a)|\leq\delta, \forall a\in \mathrm{X}} |\mathrm{T}^n_Q| \tag{73}$$

$$\leq \frac{1}{n} log \binom{n + |\mathrm{X}| - 1}{n} + \max_{Q:|Q(a)-P_X(a)|\leq\delta, \forall a\in \mathrm{X}} \frac{1}{n} log |\mathrm{T}^n_Q| \tag{74}$$

$$= H(Q) \leq H(P_X) + \epsilon, \quad n \rightarrow \infty \tag{75}$$

and the lower bound can be done by the same way:

$$\frac{1}{n}log|\mathrm{T}^n_{[Q]_{\delta_\epsilon}}| \tag{76}$$

$$=\frac{1}{n}log \sum_{Q:|Q(a)-P_X(a)|\leq\delta_\epsilon,\forall a\in \mathrm{X}} |\mathrm{T}^n_Q| \tag{77}$$

$$\geq \min_{Q:|Q(a)-P_X(a)|\leq\delta_\epsilon,\forall a\in \mathrm{X}} \frac{1}{n}log|\mathrm{T}^n_Q| \tag{78}$$

$$=H(Q)\geq H(P_X)-\epsilon, \quad n\to\infty \tag{79}$$

#

Theorem 11: (Strong AEP II: Probability of Strongly Typical Set):

$$\frac{1}{n}\log Pr\{X^n \in \mathrm{T}^{nC}_{[X]_{\delta_\epsilon}}\} \leq -\delta_\epsilon \tag{80}$$

As $\lim_{n\to\infty}\delta_\epsilon = 0$, the probability of Strongly typical set tends to 1. #

Proof: for all types $\mathrm{T}^n_Q \subseteq \mathrm{T}^{nC}_{[Q]_{\delta_\epsilon}}$, $\exists a \in \mathrm{X}, |Q(a)-P_X(a)| \geq \delta_\epsilon$. Due to the continuity and monotony of $f(p)=logp$, we have:

- Note: (here follows from the definition of derivatives, and is not accurate. In fact, $(logP_X(a)-logQ(a)) = \frac{d(logp)}{dp}|_{p=Q(a)}[P_X(a)-Q(a)]+o((P_X(a)-Q(a))^2)$is the Taylor formula with Peano remainder and the results should be $\delta_\epsilon + Q(a)o(\delta^2)$. Here we omit the $o(\delta^2)$, and we will configure δin (here) to make $\lim_{n\to\infty} nQ(a)o(\delta^2)=\infty\}$

$$D(Q||p) = |H(p,Q)-H(Q)| \tag{81}$$

$$=|\sum_{a\in \mathrm{X}} Q(a)(\log p(a)-\log Q(a))| \tag{82}$$

$$\geq Q(a)|\log p(a)-\log Q(a)| \tag{83}$$

$$=Q(a)\frac{d(logp)}{dp}|_{p=Q(a)}|P_X(a)-Q(a)| < aid = "taylor" > 8 < /a > \tag{84}$$

$$\geq Q(a)\frac{1}{Q(a)}\delta_\epsilon = \delta_\epsilon > 0 \tag{85}$$

Then we give the upper bound of atypical set:

$$\frac{1}{n}logPr(X^n \in \mathrm{T}_{[Q]_{\delta_\epsilon}}^{nC}) \tag{86}$$

$$=\frac{1}{n}log \sum_{Q:\exists a\in \mathrm{X},|Q(a)-p(a)|\geq\delta_\epsilon} |\mathrm{T}_Q^n| \tag{87}$$

$$\leq\frac{1}{n}log(\begin{matrix} n+|\mathrm{X}|-1 \\ n \end{matrix}) + \max_{Q:|Q(a)-p(a)|\leq\delta,\forall a\in \mathrm{X}} \frac{1}{n}logPr(X^n \in \mathrm{T}_Q^n) \tag{88}$$

$$=-D(Q||p)\leq -\delta_\epsilon, <aid="proaty">9 \tag{89}$$

#

Theorem 12: (Strong AEP III: The probability of Strongly Typical Sequence):

$$\forall\delta>0,\exists\eta>0 s.t. -H(X)-\eta\leq\frac{1}{n}\log p(x^n)\leq -H(X)+\eta \tag{90}$$

#

Proof: From AEP I and AEP II, we have

$$-H(p)-\epsilon-\delta\leq\frac{1}{n}\log p(x^n)=-H(p,Q)=-H(Q)-D(Q||p)\leq -H(p)+\epsilon \tag{91}$$

and letting $\eta=\delta+\epsilon$. #

One to be noticed is that in order to get a proper $\epsilon$to make both ϵitself and $n\delta^2$be sufficiently small and $n\delta_\epsilon$sufficiently large. For example,

$$\epsilon=n^p,\quad p\in(-1,-\frac{1}{2}], <aid="make">10 \tag{92}$$

is a proper setting.

Substantial Set

Definition 12: (Substantial Set):\A set $\mathrm{A}\in\mathrm{X}^n$is called a substantial set iff

$$Pr(X^n\in \mathrm{A})=\mu>0 \tag{93}$$

#

and intuitively, the positive probability can only be valued in the typical set when $n\longrightarrow\infty$, as (here) shows that :

$$Pr(X^n\in \mathrm{T}_{[Q]_{\delta_\epsilon}}^{nC})\rightarrow 0,\quad n\rightarrow\infty \tag{94}$$

To show this, we first calculate the probability of intersection of $\mathrm{A}$and typical set:

$$
\begin{aligned}
& Pr(X^n \in \mathrm{A} \cap \mathrm{T}^n_{[Q]_\delta}) && (95)\\
=& 1 - Pr(\overline{\mathrm{A}} \cup \mathrm{T}^{nC}_{[Q]_{\delta_\epsilon}}) && (96)\\
\geq& 1 - Pr(X^n \in \overline{\mathrm{A}}) - Pr(X^n \in \mathrm{T}^{nC}_{[Q]_{\delta_\epsilon}}) && (97)\\
\geq& 1 - (1-\mu) - exp^{-n\delta_\epsilon} \to \mu, \quad n \to \infty && (98)
\end{aligned}
$$

so the probability almostly distributed in the typical set. We can calculate a lower bound of $\mathrm{A}$by the property that the distribution in a typical set is uniform:

$$
\begin{aligned}
\frac{1}{n} log|\mathrm{A}| \geq& \frac{1}{n} log|\mathrm{A} \cap \mathrm{T}^n_{[Q]_\delta}| && (99)\\
\geq& \frac{1}{n} log(\mu - exp(-n\delta)) + H(X) - \epsilon && (100)\\
\to& H(X) - \epsilon, \quad n \to \infty && (101)
\end{aligned}
$$

So any non-zero part of strongly typical set has the same exponential approximation as the strongly typical set itself.

Stong typicality & Weak typicality

The strong typicality is stronger in the sense that is can imply the weak typicality.

Theorem 13: $\forall \delta > 0, \exists \eta > 0$, s.t. $\mathrm{T}^n_{[X]_\delta} \subseteq \mathrm{W}^n_{[X]_\eta}$ # The converse is not true. A counter example is that for $X \sim p = (1/2, 1/4, 1/4)$, when $x^n$is a weakly typical sequence, then we need:

$$
-\frac{1}{n} \log p(x) = -q(0) \log 1/2 - q(1) \log 1/4 - q(2) \log 1/4 \approx H(X) \qquad (102)
$$

by letting $q(0) = q(1) = 0.5, q(2) = 0$, the sequence (without realization of $p(2)$) is weekly typical, but apparently not strongly typical.

Stronger as the strong typicality is, it can only be used for RV under finite alphabets while the weak one corresponds to the weak LLN. Unless specified, we always use the term "typicality" to denote the strong typicality.

Joint Typicality

Now we come into bivariate distribution scenario where 2 generic RVs $X \in \mathrm{X}, Y \in \mathrm{Y}$ are taken into consideration to produce the i.i.d. information source $\{X_k, Y_k\}_{k\in\mathrm{N}}$.

Properties of Jointly Typical Set

Definition 13: (Jointly typical set):\ Let $\{X_k, Y_k\}_{k\in\mathrm{N}}$ be an i.i.d. information source with 2 generic RVs $X \in \mathrm{X}, Y \in \mathrm{Y}$ under finite alphabets. The jointly typical set $\mathrm{T}^n_{[XY]_\delta}$ is

$$\mathrm{T}^n_{[XY]_\delta} = \{(x^n, y^n) \in \mathrm{X}^n \times \mathrm{Y}^n : N(x, y; x^n, y^n) = 0, \forall (x, y) \notin S_{X,Y} \wedge \tag{103}$$

$$\left| \frac{N(x, y; x^n, y^n)}{n} - p(x, y) \right| \leq \delta, \forall (x, y) \in \mathrm{X} \times \mathrm{Y} \} \tag{104}$$

$$T^n_{[XY]_\delta} = \{(x^n, y^n) \in \mathrm{X}^n \times \mathrm{Y}^n : N(x, y; x^n, y^n) = 0, \forall (x, y) \notin S_{X,Y} \wedge \tag{105}$$

$$\sum_{x\in\mathrm{X}} \sum_{y\in\mathrm{Y}} \left| \frac{N(x, y; x^n, y^n)}{n} - p(x, y) \right| \leq \delta \} \tag{106}$$

$\forall (x^n, y^n) \in \mathrm{T}^n_{[XY]_\delta}$ is called jointly typical. #

Theorem 14:(Properities of Jointly Typical Set): * (Consistency): The jointly typical implies the typical separately:

$$\forall \delta > 0, \exists \delta_1, \delta_2 > 0, \lim_{n\to\infty} (\delta, \delta_1, \delta_2) = 0, s.t. \quad \mathrm{T}_{[XY]_\delta} \subseteq \mathrm{T}^n_{[X]_\delta} \times \mathrm{T}^n_{[Y]_\delta} \tag{107}$$

- (Preservation): Let $Y = f(X)$, then $x^n = (x_1, \cdots, x_n) \in \mathrm{T}^n_{[X]_\delta}$ implies $(f(x_1), \cdots, f(x_n)) \in \mathrm{T}^n_{[Y]_\delta}$

#

Proof: The consistency of JT set can be proved by noticing $N(x; x^n) = \sum_{y\in\mathrm{Y}} N(x, y; x^n, y^n)$

$$\left| \frac{N(x; x^n)}{n} - p(x) \right| = \left| \sum_{y\in\mathrm{Y}} \left(\frac{N(x, y; x^n, y^n)}{n} - p(x, y) \right) \right| \leq |\mathrm{Y}| \delta = \delta_1 \tag{108}$$

So $\forall (x^n, y^n) \in \mathrm{T}^n_{[XY]_\delta}$, then $x^n \in \mathrm{T}^n_{[X]_\delta}$. Similarly,$\forall (x^n, y^n) \in \mathrm{T}^n_{[XY]_\delta} \to y^n \in \mathrm{T}^n_{[Y]_\delta}$.

For preservation, we notice that $N(y, (f(x_1), \cdots, f(x_n))) = \sum_{x\in f^{-1}[y]} N(x; x^n)$.

$$\left|\frac{N(y;y^n)}{n}-p(y)\right| = \left|\sum_{x\in f^{-1}[y]}\frac{N(x;x^n)}{n}-p(x)\right| \tag{109}$$

$$\leq \sum_{x\in f^{-1}[y]}\left|\frac{N(x;x^n)}{n}-p(x)\right| \leq |f^{-1}[y]|\delta = \delta' \tag{110}$$

Therefore, $y^n = (f(x_1),\cdots,f(x_n)) \in \mathrm{T}^n_{[Y]_{\delta'}}$. #

Joint AEP

Definition 14: Let $(X^n, Y^n) = ((X_1, Y_1),\cdots,(X_n, Y_n))$be i.i.d. sequence with generic RV (X, Y), then there exists $\epsilon,\eta > 0$s.t. $\epsilon,\eta \to 0$as $\delta \to 0$, and

$$\forall (x^n,y^n) \in \mathrm{T}^n_{[XY]_\delta}, \quad -H(X,Y)-\epsilon \leq \frac{1}{n}\log p(x^n,y^n) \leq -H(X,Y)+\epsilon \tag{111}$$

$$\frac{1}{n}\log Pr\{(X^n,Y^n)\in \mathrm{T}^n_{[XY]_\delta}\} \leq -\delta \tag{112}$$

$$H(X,Y)-\eta \leq \frac{1}{n}\log|\mathrm{T}^n_{[XY]_\delta}| \leq H(X,Y)+\eta \tag{113}$$

#

Proof: The proof is similar to the proof of strong AEP, omitted. #

Conditional Typicality

From the JAEP, we know that when $n\to\infty$, $\frac{1}{n}\log|\mathrm{T}^n_{[XY]_\delta}| \to H(X,Y)$and $\frac{1}{n}\log|\mathrm{T}^n_{[X]_\delta}| \to H(X)$, and that (X^n,Y^n)(resp. X^n) is uniformly distributed in $\mathrm{T}^n_{[XY]_\delta}$(resp. $\mathrm{T}^n_{[X]_\delta}$). Then for each $x^n \in \mathrm{T}^n_{[X]_\delta}$, the number of typical sequence $(x^n,y^n)$are exponentially the same, each with $\frac{1}{n}\log\frac{|\mathrm{T}^n_{[XY]_\delta}|}{|\mathrm{T}^n_{[X]_\delta}|} \to H(Y|X)$. Thus we can define the typicality of y^n conditioning on a given $x^n \in \mathrm{T}^n_{[X]_\delta}$.

Definition 15: (Conditional Typical Set):\ For any $x^n \in \mathrm{T}^n_{[X]_\delta}$, the sequence $y^n \in \mathrm{T}^n_{[Y]_\delta}$ which makes (x^n,y^n)jointly typical are called the typical sequence conditioning on x^n, i.e.

$$\mathrm{T}^n_{[Y|X]_\delta}(x^n) = \{y^n \in \mathrm{T}^n_{[Y]_\delta} : (x^n,y^n)\in \mathrm{T}^n_{[XY]_\delta}\}. \tag{114}$$

The $\mathrm{T}^n_{[Y|X]_\delta}(x^n)$is called the typical set of Y conditioning on x^n. #

Theorem 15: (Conditional AEP):

$$\forall x^n \in \mathrm{T}^n_{[X]_\delta} \text{ s.t.} |\mathrm{T}^n_{[Y|X]_\delta}(x^n)| \geq 1, \exists \eta > 0, \eta \to 0 \text{ as } \delta \to 0 \text{ we have:} \quad (115)$$

$$H(Y|X) - \eta \leq |\mathrm{T}^n_{[Y|X]_\delta}(x^n)| \leq H(Y|X) + \eta \quad (116)$$

#

Proof: $\forall x^n \in \mathrm{T}^n_{[X]_\delta}$, we have

$$p(x^n) = \sum_{y^n \in \mathrm{Y}^n} p(x^n, y^n) \quad (117)$$

$$\geq \sum_{y^n \in \mathrm{T}^n_{[Y|X]_\delta}(x^n)} p(x^n, y^n) = |\mathrm{T}^n_{[Y|X]_\delta}(x^n)| p(x^n, y^n), \quad \forall y^n \in \mathrm{T}^n_{[Y|X]_\delta}(x^n) \quad (118)$$

so the upper bound of the size can be derived by the bounds from JAEP and AEP:

$$\frac{1}{n} \log |\mathrm{T}^n_{[Y|X]_\delta}(x^n)| \leq \frac{1}{n} \log(p(x^n) - p(x^n, y^n)) \quad (119)$$

$$\leq -H(X) + \eta_1 - (-H(X, Y) - \eta_2) = H(Y|X) + \eta, \quad \eta = \eta_1 + \eta_2 \quad (120)$$

Actually, $\forall (x^n, y^n) \in \mathrm{T}^n_{[XY]_\delta}$ with emperical distribution K, the size of the type with K is:

$$|\mathrm{T}^n_K| = \frac{n!}{\prod_{(x,Y) \in \mathrm{X} \times \mathrm{Y}} (nK(x, y))!} = \frac{n!}{\prod_{(x,Y) \in \mathrm{X} \times \mathrm{Y}} N(x, y; x^n, y^n)!} \quad (121)$$

for fixed $x^n \in \mathrm{T}^n_{[X]_\delta}$, the size reduces to:

$$\prod_{x \in \mathrm{X}} \frac{N(x; x^n)!}{\prod_{y \in \mathrm{Y}} N(x, y; x^n, y^n)!} \quad (122)$$

which can be a lower bound of the $|\mathrm{T}^n_{[XY]_\delta}(x^n)|$. i.e.,

$$\frac{1}{n}\log|\mathrm{T}^{n}_{[Y|X]_\delta}(x^n)| \tag{123}$$

$$\geq \frac{1}{n}\log\prod_{x\in\mathrm{X}}\frac{N(x;x^n)!}{\prod_{y\in\mathrm{Y}}N(x,y;x^n,y^n)!} \tag{124}$$

$$\overset{a}{\geq}\frac{1}{n}\sum_{x\in\mathrm{X}}(N(x;x^n)\log N(x;x^n)-N(x;x^n)) \tag{125}$$

$$-\frac{1}{n}\sum_{x\in\mathrm{X}}\sum_{y\in\mathrm{Y}}((N(x,y;x^n,y^n)+1)\log(N(x,y;x^n,y^n)+1)-N(x,y;x^n,y^n)) \tag{126}$$

$$=\frac{1}{n}\sum_{x\in\mathrm{X}}\left(N(x;x^n)\log N(x;x^n)-\sum_{y\in\mathrm{Y}}(N(x,y;x^n,y^n)+1)\log(N(x,y;x^n,y^n)+1)\right) \tag{127}$$

$$\overset{b}{\geq}\frac{1}{n}\sum_{x\in\mathrm{X}}\left(N(x;x^n)\log(n(p(x)-\delta))-\sum_{y\in\mathrm{Y}}(N(x,y;x^n,y^n)+1)\log(n(p(x,y)+\delta+\frac{1}{n}))\right) \tag{128}$$

$$=\sum_{x\in\mathrm{X}}\left(\frac{N(x;x^n)}{n}\log(p(x)-\delta)-\sum_{y\in\mathrm{Y}}(\frac{N(x,y;x^n,y^n)+1}{n})\log(p(x,y)+\delta+\frac{1}{n})\right) \tag{129}$$

$$+\frac{1}{n}\sum_{x\in\mathrm{X}}\left(N(x;x^n)-\sum_{y\in\mathrm{Y}}(N(x,y;x^n,y^n)+1)\right)\log n \tag{130}$$

$$\overset{c}{\geq}\sum_{x\in\mathrm{X}}\left((p(x)+\delta)\log(p(x)-\delta)-\sum_{y\in\mathrm{Y}}(p(x,y)-\delta+\frac{1}{n})\log(p(x,y)+\delta+\frac{1}{n})\right)-\frac{\log n}{n}|\mathrm{X}||\mathrm{Y}| \tag{131}$$

$$=H(Y|X)+\delta\sum_{x\in\mathrm{X}}(\log p(x)-1)-\delta\sum_{(x,y)\in\mathrm{X}\times\mathrm{Y}}(1-\log p(x,y)-\frac{2}{np(x,y)}) \tag{132}$$

$$-\frac{1}{n}\left(\sum_{(x,y)\in\mathrm{X}\times\mathrm{Y}}(\log p(x,y)+1)+\log n|\mathrm{X}||\mathrm{Y}|\right) \tag{133}$$

$$\leq H(Y|X)-\eta(n,\delta) \tag{134}$$

where (a) is the Strling's approximation:

$$n\ln n-n<\ln n!<(n+1)\ln(n+1)-n \tag{135}$$

and (b), (c) are from the AEP and JAEP, $\lim_{n\to\infty}\eta(n,\delta)=0$. Thus the bounds are proved. #

- Note: (c)'s bound should be changed into $(p(x,y)+\delta+\frac{1}{n})$when $\exists x,y,\quad p(x,y)=1$ because in this case $\log(p(x,y)+\delta+\frac{1}{n})$is positive. The results are the same.}

The CAEP shows that the rate of size of $|\mathrm{T}^n_{[XY]_\delta}(x^n)|$approximates $H(Y|X)$regardless of $x^n \in \mathrm{T}^n_{[X]_\delta}$. As a corollary, we show that such typical x^nthat makes $\frac{1}{n}|\mathrm{T}_{[Y|X]^n_\delta}(x^n)| \to H(Y|X)$grows with $n$at almost the same rate as the number of typical $x^n \in \mathrm{T}^n_{[X]_\delta}$

Lemma 11: Let

$$\mathrm{S}^n_{[X]_\delta} = \{x^n \in \mathrm{T}^n_{[X]_\delta} : \mathrm{T}^n_{[Y|X]_\delta}(x^n) = \emptyset\} \tag{136}$$

then

$$\frac{1}{n}\log|\mathrm{S}^n_{[X]_\delta}| \to H(X), \quad n \to \infty \tag{137}$$

$$\exists \gamma(n) \to 0, \quad \frac{1}{n}\log Pr\{X^n \in \mathrm{S}^{nC}_{[X]_\delta}\} < \gamma \tag{138}$$

#

- Note: From this lemma, we can make confusion on the $\mathrm{S}^n_{[X]_\delta}$and $\mathrm{T}^n_{[X]_\delta}$as they have the same asymptotic property in exponential sense.} **Proof:** The proof is intuitively and techniquely obvious, omitted. # The JAEP and CAEP for 3 or more RVs is similarly defined and proved, we skip the relation.

The relations of Typicality and SHannon's Inequalities

The typicality gives us an asymptotic perspective on the meaning of Shannon's measures. Correspondingly, it is highly related to the Basic inequalities.

Example 4: Consider 3 Rvs X, Y, Z.

$$(x^n, y^n, z^n) \in \mathrm{T}_{[XYZ]^n_\delta} \tag{139}$$

$$\overset{\text{Consistency}}{\Longrightarrow} (x^n, z^n) \in \mathrm{T}_{[XZ]^n_\delta}, \quad (y^n, z^n) \in \mathrm{T}_{[YZ]^n_\delta} \tag{140}$$

$$\overset{\text{CTypical}}{\Longrightarrow} (x^n) \in \mathrm{T}_{[X|Z]^n_\delta}(z^n), \quad (y^n) \in \mathrm{T}_{[Y|Z]^n_\delta}(z^n) \tag{141}$$

$$\Longrightarrow \mathrm{T}_{[XY|Z]^n_\delta}(z^n) \subseteq \mathrm{T}_{[X|Z]^n_\delta}(z^n) \times \mathrm{T}_{[Y|Z]^n_\delta}(z^n) \tag{142}$$

$$\Longrightarrow |\mathrm{T}_{[XY|Z]^n_\delta}(z^n)| \le |\mathrm{T}_{[X|Z]^n_\delta}(z^n)||\mathrm{T}_{[Y|Z]^n_\delta}(z^n)| \tag{143}$$

$$\Longrightarrow \frac{1}{n}\log|\mathrm{T}_{[XY|Z]^n_\delta}(z^n)| \le \frac{1}{n}\log|\mathrm{T}_{[X|Z]^n_\delta}(z^n)| + \frac{1}{n}\log|\mathrm{T}_{[Y|Z]^n_\delta}(z^n)| \tag{144}$$

$$\Longrightarrow H(X,Y|Z) - \epsilon_1 \le H(X|Z) + \epsilon_2 + H(Y|Z) + \epsilon_3 \tag{145}$$

$$\overset{n\to\infty,\epsilon_1,\epsilon_2,\epsilon_3\to 0}{\Longrightarrow} H(X,Y|Z) \le H(X|Z) + H(Y|Z) \tag{146}$$

A customized idea is summarized as:

- write the inequality in the form of (conditional) entropy;
- Use the jointly typical set and the property;
- Define the conditional Joint typical set;
- Give the including relation of the typical sets;
- Use AEP to transform the including relation of sets to the inequality relation of rates;
- limit $n$to infinity, get the corresponding Shannon's Inequalities.