2020 (1)

$\log$ DENOTES LOGARITHM TO THE BASE 2

$\ln$ DENOTES LOGARITHM TO THE BASE e

ELEMENTARY STATEMENT

$$\log x \le \frac{x-1}{\ln 2} \quad (*)$$

FOR ALL $x>0$, EQUALITY HOLDS IFF $x=1$.

PROOF WE CONSIDER $f(x) = \log x - \frac{x-1}{\ln 2}$

$f'(x) = \frac{1}{x \ln 2} - \frac{1}{\ln 2}$: IT EQUALS 0 AT $x=1$

$f''(x) = -\frac{1}{x^2 \ln 2} < 0$: $f(x)$ IS STRICTLY CONCAVE HENCE $x=1$ IS THE UNIQUE GLOBAL MAXIMUM

LEMMA (LOG-SUM INEQUALITY) FOR ARBITRARY NONNEGATIVE NUMBERS a_i, b_i, $i=1,\dots,n$ WE HAVE

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \ge a \cdot \log \frac{a}{b},$$

WHERE $a = \sum_{i=1}^{n} a_i$, $b = \sum_{i=1}^{n} b_i$.

THE EQUALITY HOLDS IFF $a_i b = b_i a$ FOR $i=1,\dots,n$

(HERE $\tau \cdot \log \frac{\tau}{0}$ IS DEFINED TO BE 0 IF $\tau = 0$ AND $+\infty$ IF $\tau > 0 = 0$: CONVENTION BASED ON CONTINUITY)

PROOF

- WE MAY ASSUME THAT THE a_i-S ARE POSITIVE SINCE DELETING THE PAIRS (a_i, b_i) WITH $a_i = 0$ (IF ANY), THE LEFT-HAND SIDE REMAINS UNCHANGED WHILE THE RIGHT-HAND SIDE DOES NOT DECREASE
- THEN WE CAN ALSO ASSUME THAT THE b_i-S ARE ALSO POSITIVE (OTHERWISE THE LEFT-HAND SIDE IS EQUAL TO $+\infty$)
- FURTHER, IT SUFFICES TO PROVE THE LEMMA FOR $a=b$, SINCE MULTIPLYING THE b_i-S BY A CONSTANT DOES NOT AFFECT THE INEQUALITY

IN THE LOGICAL CASE SUBSTITUTING $x = \frac{b_i}{a_i}$ IN (X) WE GET THAT

$$-\sum_{i=1}^{n} a_i \log \frac{b_i}{a_i} \geq -\sum_{i=1}^{n} a_i \frac{\left(\frac{b_i}{a_i} - 1\right)}{\ln 2} = -\frac{\sum_{i=1}^{n} b_i + \sum_{i=1}^{n} a_i}{\ln 2} = 0$$

$A = \{a_1, a_2, \dots, a_{|A|}\}$: FINITE SET OF CARDINALITY $|A|$

LET $P = \{P(a), a \in A\}$ AND $Q = \{Q(a), a \in A\}$ BE PROB. DISTRIBUTIONS ON THE SET A. THEY HOLD EACH - LEIBLER DIVERGENCE (INFORMATION DIVERGENCE, I-DIVERGENCE, RELATIVE ENTROPY) IS DEFINED BY

$$D(P\|Q) = \sum_{a:\, P(a)\neq 0} P(a) \log \frac{P(a)}{Q(a)}$$ (DUE TO THE CONVENTION IT IS

FINITE IFF $P \ll Q$, i.e., IF $Q(a) = 0$ IMPLIES $P(a) = 0$)

HOW MUCH WE CAN DISTINGUISH P AND Q WITH STATISTICAL TEST

LOG-SUM INEQUALITY IMPLIES THAT $D(P\|Q) \geq 0$ WITH EQUALITY IFF $P = Q$.

CONVERGENCE OF PROB. DISTRIBUTIONS $P_n \to P$ MEANS POINTWISE CONVERGENCE, THAT IS, $P_n(a) \to P(a)$ FOR EACH $a \in A$. TOPOLOGICAL CONCEPTS FOR PROB. DISTRIBUTIONS (CONTINUITY, OPEN AND CLOSED SETS, ETC.) ARE MEANT FOR THE TOPOLOGY INDUCED BY POINTWISE CONVERGENCE

KL DIVERGENCE IS JOINTLY SEMI CONTINUOUS IN THE PAIR (P, Q), i.e., $(P_n, Q_n) \xrightarrow{n\to\infty} (P, Q)$ IMPLIES THAT $\liminf_{n\to\infty} D(P_n\|Q_n) \geq D(P\|Q)$.

IT IS CONTINUOUS AT EACH (P, Q) WITH STRICTLY POSITIVE Q.

PROPOSITION $D(P\|Q)$ IS A CONVEX FUNCTION OF THE PAIR (P, Q)

PROOF SUPPOSE THAT $P = \alpha \cdot P_1 + (1-\alpha) P_2$, $Q = \alpha \cdot Q_1 + (1-\alpha) Q_2$, i.e., $P(a) = \alpha \cdot P_1(a) + (1-\alpha) P_2(a)$ FOR EVERY $a \in A$ AND SIMILARLY FOR Q, WHERE $0 < \alpha < 1$.

LOG-SUM INEQUALITY IMPLIES THAT FOR ALL $a \in A$

$$\alpha \cdot P_1(a) \cdot \log \frac{P_1(a)}{Q_1(a)} + (1-\alpha) \cdot P_2(a) \cdot \log \frac{P_2(a)}{Q_2(a)} =$$

$$= \alpha \cdot P_1(a) \cdot \log \frac{\alpha P_1(a)}{\alpha Q_1(a)} + (1-\alpha) \cdot P_2(a) \log \frac{(1-\alpha) P_2(a)}{(1-\alpha) Q_2(a)} \geq P(a) \cdot \log \frac{P(a)}{Q(a)}$$

SUMMING FOR $a \in A$ GIVES $\alpha D(P_1 \| Q_1) + (1-\alpha) \cdot D(P_2 \| Q_2) \geq D(P \| Q)$ □

REMARK: THE ABOVE PROOF SHOWS ALSO THAT $D(P\|Q)$ IS STRICTLY CONVEX IN P IF $Q(a) > 0$ FOR ALL a.

VARIATIONAL DISTANCE (SOME AUTHORS USE THE TERM FOR THE HALF OF THIS):

$$\|P - Q\| = \sum_a |P(a) - Q(a)|$$

PINSKER INEQUALITY (ALSO CALLED CSISZÁR–KEMPERMAN–KULLBACK–PINSKER INEQ.)

$$D(P\|Q) \geq \frac{1}{2 \ln 2} \|P - Q\|^2$$

NON UNIFORM CONTINUITY: IF $D(P\|Q) < \infty$ THEN

$$D(P\|Q) \leq \frac{\|P-Q\|^2}{\alpha_Q} \quad \text{where } \alpha_Q = \min_{a: Q(a) \neq 0} Q(a)$$

NUMBERS WITH APPROPRIATE UNITS

THE ENTROPY OF A RAND. VAR. $P = \{P(a), a \in A\}$ IS DEFINED BY

$$H(P) = -\sum_{a \in A} P(a) \log P(a)$$

$H(P) \geq 0$ BY DEFINITION

(*) IS AN EQUALITY IF $\forall a \in A$

$\alpha \cdot P_1(a) \cdot (\alpha \cdot Q_1(a) + (1-\alpha) Q_2(a)) = \alpha \cdot Q_1(a) (\alpha \cdot P_1(a) + (1-\alpha) P_2(a))$

AND $(1-\alpha) P_2(a) (\dots) = (1-\alpha)(Q_2(a))(\dots)$

$\Leftrightarrow P_1(a) = P_2(a)$

IF $Q_1 = Q_2 = Q$ WE HAVE EQUALITY ALWAYS

$$P_1(a) = \alpha \cdot P_1(a) + (1-\alpha) P_2(a)$$
$$P_2(a) = \alpha \cdot P_1(a) + (1-\alpha) P_2(a)$$

$D(P \,||\, \text{UNIFORM DIST ON } A) = \sum_{a \in A} P(a) \cdot \log \frac{P(a)}{\frac{1}{|A|}} =$

$= \sum_{a \in A} P(a) \cdot \log P(a) + \sum_{a \in A} p(a) \log |A| = \log |A| - H(P)$

HENCE $\boxed{H(P) = \log |A| - D(P \,||\, \text{UNIFORM ON } A)}$

□ INDICES: • $H(P) \leq \log |A|$, $=$ HOLDS IFF P IS UNIFORM DIST

• $H(P)$: CONTINUOUS AND STRICTLY CONCAVE IN P

LET X BE RV TAKING VALUES IN A

$H(X) = H(P_X) = E(-\log P_X(X))$

MEASURE OF INFORMATION CONTENT OR MEASURE OF AVERAGE UNCERTAINTY ABOUT THE OUTCOME OF X

SEVERAL AXIOMATIC AND OPERATIONAL JUSTIFICATION EXIST FOR □

A^∞ DENOTES THE SET OF ALL INFINITE SEQUENCE $x = x_1^\infty$ WITH $x_i \in A$, $i \geq 1$.

A^* DENOTES THE SET OF ALL FINITE SEQUENCE DRAWN FROM A (THE EMPTY SET IS ALSO INCLUDED)

THE CONCATENATION OF $u \in A^*$ AND $v \in A^* \cup A^\infty$ IS DENOTED BY uv

A FINITE SEQUENCE u IS A PREFIX OF A FINITE OR INFINITE SEQUENCE w, AND WE WRITE $u \prec w$ IF $w = uv$ FOR SOME v

---

DATA COMPRESSION

EXAMPLE LET X BE A RV TAKING VALUES IN $A = \{1,2,3,4\}$ HAVING THE FOLLOWING DISTRIBUTION:

$$P_X(X=1) = \frac{1}{2},\ P_X(X=2) = \frac{1}{4},\ P_X(X=3) = \frac{1}{8},\ P_X(X=4) = \frac{1}{8}$$

WE WOULD LIKE TO CONVEY EFFICIENTLY ITS OUTCOME

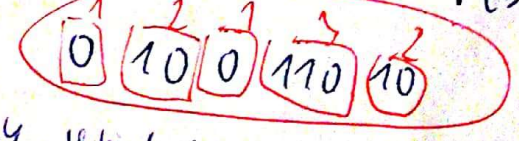
IDEA ASSIGNING SHORT DESCRIPTIONS TO THE MOST FREQUENT OUTCOMES AND NECESSARILY LONGER DESCRIPTIONS TO THE LESS FREQUENT SYMBOLS

NO CODEWORD IS A PREFIX OF ANOTHER ONE

$$C(1) = 0,\ C(2) = 10,\ C(3) = 110,\ C(4) = 111$$

THEN $E(L) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 = 1{,}75 = H(X)$

ADVANTAGE OF THE PREFIX CODE: IF WE USE THE CODE RECURSIVELY WE CAN DECODE THE SYMBOLS UNIQUELY AND INSTANTANEOUSLY (THE SYMBOL a CAN BE DECODED AS SOON AS WE COME TO THE END OF THE CODEWORD CORRESPONDING TO IT)



A CODE FOR SYMBOLS IN A WITH IMAGE ALPHABET B IS A MAPPING $C: A \to B^*$, ITS LENGTH FUNCTION $L: A \to \mathbb{N}$ IS DEFINED BY THE FORMULA $|C(a)| = L(a)$ (IN THE TEXT)

WE ASSUME: $B = \{0,1\}$, THE CODEWORDS ARE DISTINCT AND DIFFERENT FROM THE EMPTY STRING.

C IS PREFIX IF $C(a)$ IS NOT AN INITIAL SEGMENT OF $C(\tilde{a})$ WHEN $a \neq \tilde{a}$ IN A

THE DYADIC REPRESENTATIONS FOR NUMBERS $y \in [0,1)$ ARE OF THE FORM $y = \sum_{k=1}^{\infty} y_k 2^{-k}$ WITH $y_k \in \{0,1\}$.

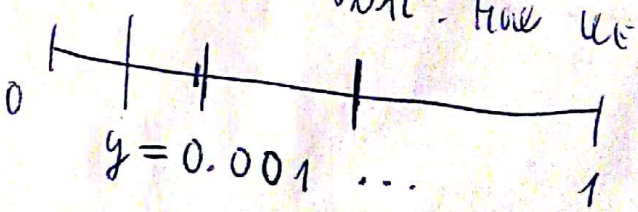
THE NUMBERS OF FORM $y = \frac{m}{2^n}$ WHERE n AND m ARE INTEGERS WITH $n \geq 1$, $0 \leq m < 2^n$, ARE CALLED DYADIC RATIONALS.

THEY HAVE TWO DYADIC REPRESENTATIONS: ONE WITH INFINITE 0-S AND ONE WITH INFINITE 1-S (E.G. $\frac{1}{4} = 0{,}01 = 0{,}00\dot{1}$).

WE WILL USE THE REPRESENTATION WITH THE INFINITE 0-S.

ALL OTHER NUMBERS IN $[0,1)$ HAVE A UNIQUE REPRESENTATION.

ASSUME THAT y IS NOT A DYADIC RATIONAL. HOW WE CAN FIND ITS REPRESENTATION



WHERE ARE THE NUMBERS WHOSE DYADIC REPRESENTATION START WITH 0 110 : IN THE INTERVAL OF LENGTH $2^{-4} = \frac{1}{16}$ FOLLOWING $0{,}0110 = 0{,}011$

LEMMA (KRAFT INEQUALITY) A FUNCTION $L: A \to \mathbb{N}$ IS THE LENGTH FUNCTION OF SOME PREFIX CODE IFF IT SATISFIES THE SO CALLED KRAFT INEQUALITY: $\sum_{a \in A} 2^{-L(a)} \leq 1$

PROOF ASSUME THAT $\ell: A \to B^*$ IS A PREFIX CODE. ASSOCIATE WITH EACH $a \in A$ THE NUMBER $t(a)$ WHOSE BINARIC IS THE CODEWORD $\ell(a) = b_1^{L(a)}$ THAT IS $t(a) = 0.b_1 \dots b_{L(a)}$

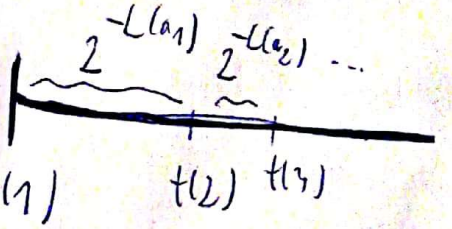
THE PREFIX CONDITION IMPLIES THAT $t(\hat{a}) \notin [t(a), t(a) + 2^{-L(a)})$ IF $\hat{a} \neq a$. HENCE THE INTERVALS $[t(a), t(a) + 2^{-L(a)})$, $a \in A$, ARE DISJOINT. THE TOTAL LENGTH OF THESE DISJOINT INTERVALS ARE $\leq$ THAN $1 \Rightarrow \sum_{a \in A} 2^{-L(a)} \leq 1$.

CONVERSELY, SUPPOSE A FUNCTION $L: A \to \mathbb{N}$ SATISFIES $\sum_{a \in A} 2^{-L(a)} \leq 1$

LABEL A SO THAT $L(a_i) \leq L(a_{i+1})$, $i = 1, \dots, |A|-1$

(OR EQUIVALENTLY $2^{-L(a_i)} \geq 2^{-L(a_{i+1})}$)

LET $t(i) = \sum_{j < i} 2^{-L(a_j)}$



$t(i)$ CAN BE BINARICALLY REPRESENTED AS $t(i) = 0.b_1 \dots b_{L(a_i)}$

THEN $\ell(a_i) = b_1^{L(a_i)}$ DEFINES A PREFIX CODE WITH LENGTH FUNCTION L.

NOTE !!! PRESCRIBED LENGTHS CAN COINCIDE !!!

KEY CONSEQUENCE OF THE LEMMA IS SHANNON'S NOISELESS CODING THEOREM

THEOREM LET P BE A PROB DIST ON A. THEN EACH PREFIX CODE HAS EXPECTED LENGTH $E(L) = \sum_{a \in A} P(a) L(a) \geq H(P)$. FURTHERMORE, THERE IS A PREFIX CODE WITH LENGTH FUNCTION $L(a) = \lceil -\log P(a) \rceil$ (AND HENCE ITS EXPECTED LENGTH SATISFIES $E(L) < H(P) + 1$)

PROOF OF THE FIRST ASSERTION

$$E(L) - H(P) = -\sum_{a \in A} P(a) \log 2^{-L(a)} + \sum_{a \in A} P(a) \log P(a) =$$

$$= \sum_{a \in A} P(a) \log \frac{P(a)}{2^{-L(a)}} \underset{\text{LOG IS}}{\geq} 1 \cdot \log \frac{1}{\sum_{a \in A} 2^{-L(a)}} \underset{\text{KRAFT}}{\geq} 0$$

PROOF OF THE SECOND ASSERTION

$L(a) = \lceil -\log P(a) \rceil$ SATISFIES THE KRAFT INEQUALITY $\nearrow \sum_{a \in A} 2^{-\lceil -\log P(a) \rceil} \leq \sum_{a \in A} 2^{+\log P(a)} \leq 1$

REMARK 1 THE ABOVE CONSTRUCTION ACHIEVING $E(L) < H(P) + 1$ IS NOT PRACTICAL WHEN $|A|$ IS LARGE, SINCE IT RELIES ON THE CONSTRUCTION USED IN THE PROOF OF THE KRAFT LEMMA ABOUT THE KRAFT INEQUALITY WHERE THE ORDERING OF THE LENGTHS IS NECESSARY

REMARK 2 IT WILL BE IMPORTANT LATER TO DISTINGUISH THE DISTRIBUTION P ACCORDING TO WHICH WE CALCULATE EXPECTED LENGTH AND THE CODING DISTRIBUTION DENOTED BY Q

ASSUME THAT $P \ll Q$. WE KNOW THAT THERE EXISTS A PREFIX CODE WITH $L(a) = \lceil -\log Q(a) \rceil$. FOR THIS CODE

$$H(P) + D(P\|Q) \leq E_P(L) \leq 1 + H(P) + D(P\|Q)$$

SINCE $\sum_{a \in A} P(a)\,(-\log Q(a)) = \sum_{a \in A} P(a) \frac{1}{\log Q(a)} + \sum_{a \in A} P(a) \log P(a)$

$$- \sum_{a \in A} P(a) \cdot \log P(a) = D(P\|Q) + H(P)$$

THE FOLLOWING THEOREM SHOWS THAT EVEN NON-PREFIX CODES CAN NOT SUBSTANTIALLY BEAT THE ENTROPY LOWER BOUNDS

THEOREM THE LENGTH FUNCTION OF A NOT NECESSARILY PREFIX CODE $C : A \to B^*$ SATISFIES $\sum_{a \in A} 2^{-L(a)} \leq \log |A|$ AND FOR ANY PROB. DIST P ON A, THE CODE HAS EXPECTED LENGTH

$$E_P(L) \geq H(P) - \log \log |A|$$

X RV TAKING VALUES IN A

$$H(X) = H(P_X) = -\sum_{a \in A} P_X(a) \log P_X(a) = E(-\log P_X(X))$$

MEASURE OF INFORMATION CONTENT, MEASURE OF AVERAGE UNCERTAINTY ABOUT THE OUTCOME OF X

LET (X,Y) BE A 2-DIM R.V. WITH VALUES IN $A \times B$

THEN $H(X,Y)$ IS ALSO WELL DEFINED

$$H(X,Y) \stackrel{\Delta}{=} -\sum_{a \in A} \sum_{b \in B} P_{XY}(a,b) \log P_{XY}(a,b) = E(-\log P_{XY}(X,Y))$$

COND ENTROPY: $H(Y|X) = \sum_{a \in A} P_X(a) \cdot H(Y|X=a) =$

$$= -\sum_{a \in A} P_X(a) \sum_{b \in B} P_{Y|X}(b|a) \log P_{Y|X}(b|a) =$$

$$= -\sum_{a \in A} \sum_{b \in B} P_{XY}(a,b) \log P_{Y|X}(b|a) = E(-\log P_{Y|X}(Y|X))$$

CHAIN RULE: $H(X,Y) = H(X) + H(Y|X)$

MUTUAL INFORMATION OF X AND Y IS DEFINED BY

$$I(X;Y) \stackrel{\Delta}{=} H(X) - H(X|Y) =$$

$$= -\sum_{a \in A} \sum_{b \in B} P_{XY}(a,b) \log P_X(a) + \sum_{a \in A} \sum_{b \in B} P_{XY}(a,b) \log P_{X|Y}(a|b) =$$

$$= \sum_{a \in A} \sum_{b \in B} P_{XY}(a,b) \cdot \log \frac{P_{X|Y}(a|b) \cdot P_Y(b)}{P_X(a) \cdot P_Y(b)} = D(P_{XY} \,||\, P_X \times P_Y)$$

WHERE $P_X \times P_Y(a,b) = P_X(a) \cdot P_Y(b)$

IT FOLLOWS THAT $I(X;Y) \geq 0$, $= 0$ IFF X AND Y ARE INDEPENDENT

NOTE ALSO THAT $I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$

IMPORTANT CONSEQUENCE: $H(X|Y) \leq H(X)$ (CONDITIONING REDUCES ENTROPY)

$$I(X;Y|Z) \stackrel{\Delta}{=} H(X|Z) - H(X|Y,Z) \geq 0$$

$$\Rightarrow H(X|Y,Z) \leq H(X|Z)$$

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

SIMILAR CALCULATIONS AS BEFORE

AT THIS POINT WE USE THE FOLLOWING DEFINITION OF STOCH. PROCESS:
IT IS AN INDEXED SEQUENCE OF RANDOM VARIABLES DEFINED ON A COMMON PROBABILITY SPACE.

WE ASSUME THAT THE INDEX SET IS $\mathbb{Z}^+$, AND THAT $X_1, X_2, \ldots$ TAKING VALUES IN A FINITE SET A

P_n DENOTES THE JOINT DIST OF $(X_1, \ldots, X_n)$: $P_n(x_1^n) = \text{Prob}\{X_1 = x_1, \ldots, X_n = x_n\}$
$x_1^n \in A^n$

NOTE THAT FOR THESE DISTRIBUTIONS THE CONSISTENCY CONDITIONS
$P_n(x_1^n) = \sum_{a \in A} P_{n+1}(x_1^n a)$ HOLDS

$X_1, X_2, \ldots X_n, \ldots$ IS STATIONARY IF $\text{Prob}\{X_1 = x_1, \ldots, X_n = x_n\} =$
$= \text{Prob}(X_{1+\ell} = x_1, \ldots, X_{n+\ell} = x_n)$ FOR EVERY n, EVERY SHIFT ℓ
AND EVERY $x_1, \ldots, x_n \in A$.

SIMPLEST WELL KNOWN EXAMPLE: I.I.D SEQUENCES

THEOREM ASSUME THAT $X_1, X_2, \ldots$ IS A STATIONARY SEQUENCE THEN BOTH $\lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}$ AND $\lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1})$ EXIST AND THEY ARE EQUAL. THE LIMIT IS CALLED THE ENTROPY RATE OF THE PROCESS, AND IT IS DENOTED BY $\bar{H}$.

LEMMA (CÉSARO MEAN): IF $a_n \to a$ AND $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$
THEN $b_n \to a$

PROOF ε FIXED, $\exists N_\varepsilon$ S.T. $|a_n - a| < \varepsilon \quad \forall n \geq N_\varepsilon$. THEN IF n IS LARGE
$|b_n - a| = \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \leq \frac{1}{n}\sum_{i=1}^{N(\varepsilon)} |a_i - a| + \frac{n - N(\varepsilon)}{n}\varepsilon \leq 2\varepsilon$

PROOF OF THE THEOREM $H(X_{n+1} | X_1, \dots, X_n) \leq H(X_{n+1} | X_2, \dots, X_n) =$

STATIONARITY

$= H(X_n | X_1, \dots, X_{n-1}) \Rightarrow \lim_{n \to \infty} H(X_n | X_1, \dots, X_{n-1})$ EXISTS

LET $\bar{H}$ DENOTE THIS LIMIT. THEN ~~$\lim_{n \to \infty} \bar{H}$~~

$$\frac{1}{n} H(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_1, \dots, X_{i-1}) \xrightarrow{n \to \infty} \bar{H}$$

CESÀRO

IN REAL-WORLD APPLICATIONS WE HAVE TO COMPRESS STOCHASTIC PROCESSES (DATA SOURCE) $X_1, X_2, \dots, X_n, \dots$. WE CAN ENCODE EACH RANDOM VARIABLE SEPARATELY BUT DUE TO THE 1-BIT OVERHEAD IT IS CLEARLY NOT THE BEST EVEN FOR I.I.D. SOURCE

$X_1 X_2 \dots X_n \dots$

n-CODES: WE CAN REDUCE THE OVERHEAD PER SYMBOL BY SPREADING IT OUT OVER MANY SYMBOLS

IN THIS CASE THE CODE C_n IS A MAPPING FROM A^n TO B^*.

SHANNON'S NOISELESS CODING THEOREM AND THE ABOVE CONSIDERATIONS IMPLY

THEOREM THE MINIMUM EXPECTED CODEWORD LENGTH PER SYMBOL SATISFIES:

$$\frac{H(X_1, \dots, X_n)}{n} \leq \frac{\min_{C_n} E(L_{C_n})}{n} \leq \frac{H(X_1, \dots, X_n)}{n} + \frac{1}{n}$$

MOREOVER, IF $X_1, \dots, X_n, \dots$ IS A STATIONARY PROCESS THEN

$$\frac{\min_{C_n} E(L_{C_n})}{n} \xrightarrow{n \to \infty} \bar{H}$$

LEMMA THE GAIN OF JOINT CODING CAN BE LARGE IF THE SYMBOLS OF THE DATA SOURCE ARE HIGHLY DEPENDENT.

PROBLEM) • THE CODING ALGORITHM PROVIDED BY SHANNON'S NOISELESS CODING THEOREM IS NOT PRACTICAL (WE HAVE TO ORDER THE LENGTHS ASSIGNED TO THE ELEMENTS OF A^n : NOW

• IT USES THE DISTRIBUTION OF THE THE DATA SOURCE : LATER

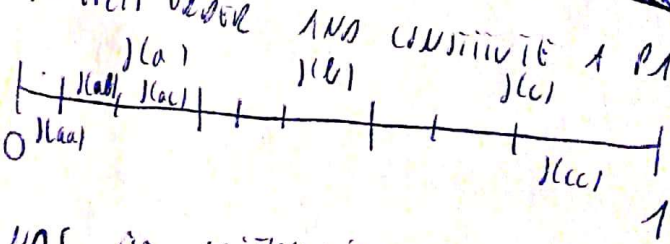
LET $Q_n, n=1,2,\dots$ BE PROBABILITY DISTRIBUTIONS ON THE SET A^n SATISFYING THE CONSISTENCY CONDITIONS $Q_n(x_1^n) = \sum_a Q_{n+1}(x_1^n a)$.

AN ARITHMETIC CODE WITH CODING DISTRIBUTION SEQUENCE $\{Q_n, n\geq 1\}$ IS A SEQUENCE OF n-CODE DEFINED AS FOLLOWS.

FIX AN ORDERING ON THE ELEMENTS OF A: $a_1 < a_2 \dots < a_{|A|}$

FOR EACH n PARTITION THE UNIT INTERVAL $[0,1)$ INTO SUBINTERVALS $J(x_1^n) = [\ell(x_1^n), r(x_1^n))$ OF LENGTH $r(x_1^n) - \ell(x_1^n) = Q_n(x_1^n)$ FOR EACH $x_1^n \in A^n$ IN A NESTED MANNER, i.e., SUCH THAT $J(x_1^n a_1), J(x_1^n a_2), \dots, J(x_1^n a_{|A|})$ FOLLOW EACH OTHER IN THIS ORDER AND CONSTITUTE A PARTITION OF $J(x_1^n)$.

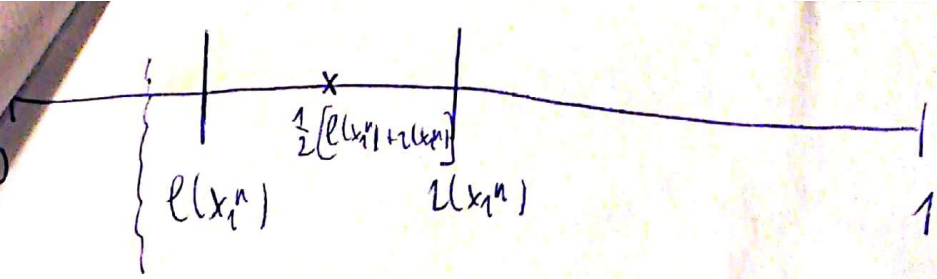
$A = \{a, b, c\}$



WE DEFINE TWO KINDS OF ARITHMETIC CODES:

1. $C(x_1^n) = z_1^m$ IF THE ENDPOINTS OF $J(x_1^n)$ HAVE BINARY EXPANSION $\ell(x_1^n) = 0.z_1 z_2 \dots z_m 0$, $r(x_1^n) = 0.z_1 z_2 \dots z_m 1$

2. $\tilde{C}(x_1^n) = z_1^{\tilde{m}}$ IF THE MIDPOINT OF $J(x_1^n)$ HAS BINARY EXPANSION $0.z_1 z_2 \dots z_{\tilde{m}}$, $\tilde{m} = \lceil -\log Q_n(x_1^n) \rceil + 1$



- $(\!(x_1^n) \prec \tilde{c}(x_1^n)$
- $L(x_1^n) < \tilde{m} = \tilde{L}(x_1^n)$
- $(\!(x_1^n)$ is 1-1 (THE CODEWORDS ARE DISTINCT)
- $\tilde{c}(x_1^n)$ is PREFIX
- (i) UNCLE: BACK $(\!(x_1^n)$ is NOT A PREFIX OF $(\!(x_1^n a)$

IN ORDER TO DETERMINE CODEWORDS $(\!(x_1^n)$ OR $\tilde{c}(x_1^n)$ THE NESTED PARTITIONS NEED NOT BE ACTUALLY COMPUTED, IT IS ENOUGH TO FIND $J(x_1^n)$. THIS CAN BE DONE IN n STEPS, THE i^{TH} STEP IS TO PARTITION THE INTERVAL $J(x_1^{i-1})$ INTO $|A|$ SUBINTERVALS OF LENGTH PROPORTIONAL TO THE CONDITIONAL PROBABILITIES

$$Q(a \mid x_1^{i-1}) = Q_i(x_1^{i-1} a) \,/\, Q_{i-1}(x_1^{i-1}), \quad a \in A$$

LET $X_1, X_2, \ldots$ BE A STOCHASTIC PROCESS AND LET P_n DENOTES THE JOINT DISTRIBUTION OF $(X_1, \ldots, X_n)$, $n \geq 1$. CONSIDER ARITHMETIC CODE WITH CODING DISTRIBUTION SEQUENCE $\{P_n, n \geq 1\}$. FOR EACH n, ITS EXPECTED CODEWORD LENGTH PER SYMBOL CAN BE UPPER-BOUNDED BY $\frac{H(X_1 \ldots X_n) + 2}{n}$. HENCE, IT IS OPTIMAL IN THE EXP. CODELENGTH P.U. n A PRACTICAL METHOD WHEN $\frac{H}{n}$ IS ACHIEVED ASYMPTOTICALLY.

LARGE DEVIATIONS VIA TYPES

THE TYPE OF A SEQUENCE $x_1^n \in A^n$ IS ITS EMPIRICAL DISTRIBUTION $\hat{P} = \hat{P}_{x_1^n}$, THAT IS THE DISTRIBUTION DEFINED BY

$$\hat{P}(a) = \frac{|\{i : x_i = a\}|}{n}, \quad a \in A$$

EXAMPLE $x_1^5 = 01001$, $\hat{P}_{x_1^5} = \left(\frac{3}{5}, \frac{2}{5}\right)$

A DISTRIBUTION P ON A IS CALLED n-TYPE IF $P = \hat{P}_{x_1^n}$ FOR SOME $x_1^n \in A^n$

$\mathcal{P}_n$ DENOTES THE SET OF n-TYPES

$A = \{0,1\}$ $\mathcal{P}_4 = \{(0,1), (\frac{1}{4},\frac{3}{4}), (\frac{1}{2},\frac{1}{2}), (\frac{3}{4},\frac{1}{4}), (1,0)\}$

T_P^n : TYPE CLASS OF n-TYPE P : IT IS THE SET OF THOSE $x_1^n \in A^n$ FOR WHICH $\hat{P}_{x_1^n} = P$.

$T^4_{(\frac{1}{4},\frac{3}{4})} = \{0111, 1011, 1101, 1110\}$

LET $A = \{a_1, \dots, a_t\}$ WHERE $t = |A|$. LET $P \in \mathcal{P}_n$ BE FIXED.

LET k_i DENOTE THE OCCURENCE OF a_i IN x_1^n FOR ANY FIXED $x_1^n \in T_P^n$ (IT DOES NOT MATTER WHICH ELEMENT OF T_P^n WE CHOOSE)

THEN THE t-TUPLE $(k_1, k_2, \dots, k_t)$ OF NONNEGATIVE INTEGERS WITH $k_1 + \dots + k_t = n$ UNIQUELY DETERMINES THE n-TYPE P.

THEN $|\mathcal{P}_n| \le (n+1)^{|A|}$ REMINDS ME OF THE FORMULA OF COMBINATION WITH REPETITION GIVES

LEMMA $$|\mathcal{P}_n| = \binom{n+|A|-1}{|A|-1}$$

LATER WE WILL USE THAT

$$\binom{n+|A|-1}{|A|-1} = \frac{(n+|A|-1)\cdots(n+1)}{(|A|-1)(|A|-2)\cdots 1} \le \frac{(n+|A|-1)}{2} \cdots \frac{n+2}{2}(n+1) \le (n+1)^{|A|-1}$$

IF $n \ge |A|-3$

IMPORTANT FACT THE NUMBER OF n-TYPES IS POLYNOMIAL IN n

LEMMA FOR ANY n-TYPE P: $\binom{n+(|A|-1)}{|A|-1}^{-1} 2^{nH(P)} \leq |T^n_P| \leq 2^{nH(P)}$

PROOF LET $(k_1, \dots, k_t)$ BE THE t-TUPLE WHICH CORRESPONDS TO THE n-TYPE P [illegible]. WE HAVE

$$|T^n_P| = \frac{n!}{k_1! k_2! \dots k_t!}$$ (PERMUTATION WITH REPETITION)

$$n^n = (k_1 + \dots + k_t)^n = (k_1 + \dots + k_t)(k_1 + \dots + k_t) \dots (k_1 + \dots + k_t) =$$

$$= \sum_{(j_1, \dots, j_t)} \frac{n!}{j_1! \dots j_t!} k_1^{j_1} \dots k_t^{j_t}$$

$(j_1, \dots, j_t)$: t-TUPLE OF NON-NEGATIVE INTEGERS WITH $j_1 + \dots + j_t = n$

NOW WE SHOW THAT THE LARGEST TERM IN [illegible] IS $\frac{n!}{k_1! \dots k_t!} k_1^{k_1} \dots k_t^{k_t}$ (*)

ASSUME THAT THIS IS NOT THE CASE. DENOTE THE LARGEST TERM BY $\frac{n!}{\ell_1! \dots \ell_t!} k_1^{\ell_1} \dots k_t^{\ell_t}$. THEN THERE EXISTS r AND s WITH $\ell_r > k_r$ AND $\ell_s < k_s$. DECREASING ℓ_r BY 1 AND INCREASING ℓ_s BY 1 MULTIPLIES THE CORRESPONDING TERM BY $\frac{\ell_r}{k_r} \frac{k_s}{\ell_s + 1} = \frac{\ell_r}{k_r} > 1$

IT GIVES A CONTRADICTION. HENCE, WE PROVED (*). THEN

VIII

$$\frac{n!}{k_1!\cdots k_t!}\,k_1^{k_1}\cdots k_t^{k_t} \le n^n = \sum_{(j_1\cdots j_t):\,t} \frac{n!}{j_1!\cdots j_t!}\,j_1^{j_1}\cdots j_t^{j_t} \le \binom{n+t-1}{t-1}\frac{n!}{k_1!\cdots k_t!}\,k_1^{k_1}\cdots k_t^{k_t}$$

TUPLE OF NONNEGATIVE INTEGERS WITH $j_1+\cdots+j_t=n$

AND

$$\frac{n^n}{k_1^{k_1}\cdots k_t^{k_t}} = \prod_{i=1}^{t}\left(\frac{k_i}{n}\right)^{-k_i} = \prod_{i=1}^{t}\left(p(a_i)\right)^{-n\,Q(a_i)} =$$

$$= 2^{\log \prod_{i=1}^{t} p(a_i)^{-n\,Q(a_i)}} = 2^{n\,H(Q)} \qquad \square$$

THE NEXT RESULT CONNECTS THE THEORY OF TYPES WITH GENERAL INFO THEORY

FOR ANY DIST P ON A LET P^n DENOTE THE DISTRIBUTION OF n INDEPENDENT RV'S FROM P, THAT IS $P^n(x_1^n) = \prod_{i=1}^{n} P(x_i)$, $x_1^n \in A^n$

LEMMA FOR ANY DISTRIBUTION P ON A AND ANY n-TYPE Q:

$$P^n(x_1^n) = 2^{-n[D(Q\|P)+H(Q)]} \quad \text{SAME FOR EACH } x_1^n \in Y_Q^n$$

$$\binom{n+|A|-1}{|A|-1}^{-1} 2^{-n\,D(Q\|P)} \le P^n(Y_Q^n) \le 2^{-n\,D(Q\|P)}$$

IF $x_1^n \in Y_Q^n$ AND

PROOF $P^n(x_1^n) = \prod_{a\in A} P(a)^{n\cdot Q(a)} = 2^{n\sum_a Q(a)\log P(a)} =$

$$= 2^{-n\sum_a Q(a)\log\frac{1}{P(a)} - n\sum_a Q(a)\log Q(a) + n\sum_a Q(a)\log Q(a)} = 2^{-n[D(Q\|P)+H(Q)]}$$

FINALLY, $P^n(Y_Q^n) = |Y_Q^n|\cdot P^n(x_1^n)$ FOR ANY FIXED $x_1^n \in Y_Q^n$.

HENCE, THE SECOND CLAIM FOLLOWS BY THE FIRST CLAIM AND THE PREVIOUS LEMMA.