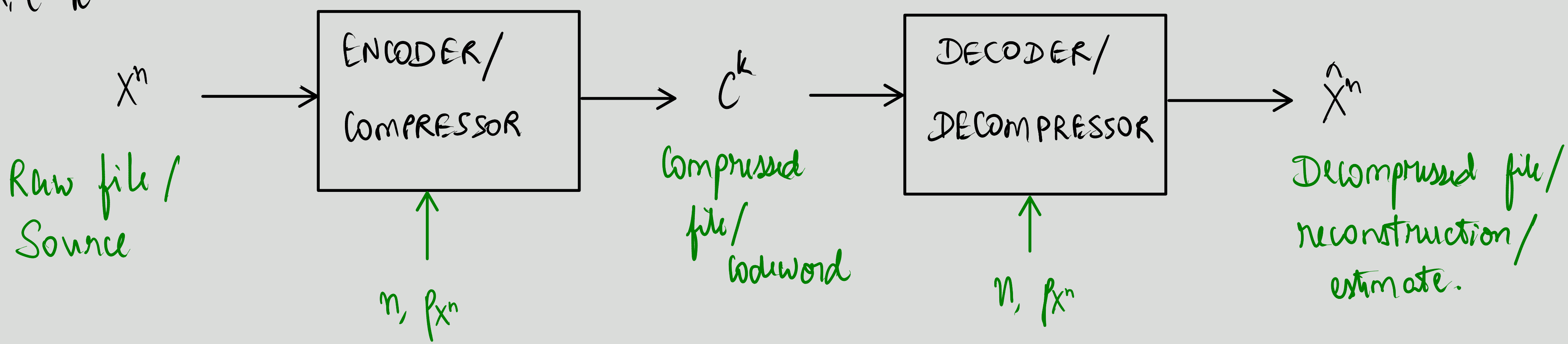
Source Coding/Data Compression

Claude Shannon, "A Mathematical Theory of Communication"

$X_1 \cdots X_n$

$X_i \in \mathcal{X}$

$c^k \in \{0,1\}^k$



p_{X^n} : Source distribution

Universal

Memoryless : $X^n \sim \text{iid}(p_X)$

Rate : $R = \dfrac{k}{n}$

Assumptions:

1. X^n is random ✓
2. $X^n \sim \text{iid}(p_X)$ ✗
3. p_X is known to ENC DEC ✗

p_{X^n}

1. Lossless Compression: $\hat{X}^n \approx X^n$

Fixed-length compression: k · depends only on n, p_{X^n}

P_e = Probability of error

Harry Potter	I	1 MB	100 kB
	II	1 MB	100 kB

$R = \frac{k}{n}$

Variable-length compression: → Expected length $= \mathbb{E}\, k(X^n)$ $\quad R_{avg} = \frac{\mathbb{E}\, k(X^n)}{n}$

$\hat{X}^n = X^n$

k can vary even if n, p_{X^n} fixed

1 MB	100 kB
1 MB	102 kB

2. Lossy Compression: $\hat{X}^n \neq X^n$

Distortion measure $\quad \mathbb{E}\, d(X^n, \hat{X}^n) < \delta$

$$MSD = \mathbb{E}\, \|\hat{X}^n - X^n\|^2 = \mathbb{E} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2$$

$n=2$

0 0	0.5	$\longrightarrow$ 0	$\longrightarrow$	0 0
0 1	0.25	$\longrightarrow$ 1	$\longrightarrow$	0 1
1 0	0.125	$\longrightarrow$ 00	$\longrightarrow$	1 0
1 1	0.125	$\longrightarrow$ 01	$\longrightarrow$	1 1

$\uparrow$ $P(x^n)$ $\quad$ C^k

2 bits

Expected length: $\mathbb{E}\, k(X^n)$

$$= 0.5\times 1 + 0.25\times 1 + 0.125\times 2 + 0.125\times 2$$

$$= 0.5 + 0.25 + 0.5$$

$$= 1.25 \text{ bits} < 2 \text{ bits.}$$

$$R_{avg} = \frac{1.25}{2} < 1$$

ENC

0 0	0.5	$\longrightarrow$	0
0 1	0.25	$\longrightarrow$	1
1 0	0.125	$\longrightarrow$	0
1 1	0.125	$\longrightarrow$	1

DEC

$0 \longrightarrow 00$

$1 \longrightarrow 01$

$k = 1$ bit

P_e

Probability of error

$$= \Pr[\hat{X}^n \neq X^n]$$

$$= p(10) + p(11) = 0.25$$

$$R_{avg} = \frac{\mathbb{E}\, k(X^n)}{n}$$

Optimal variable-length compression

Assumptions: (n, P_{X^n}) is known,

Computational power - free.

Want $\mathbb{E}\, k(X^n)$ to be as small as possible.

$P_{X^n}(x^n) \uparrow \Rightarrow k(X^n) \downarrow$

$X^n \in \mathcal{X}^n$ $\quad \mathcal{X} = \{0, 1, 2 \cdots a\}$

$x_1 \; - - - \; x_n$

C^k

$$(a+1)^n \left\{ \begin{array}{ll} x^n(1) & P_{X^n}(x^n(1)) \\ & \quad \vee\!/ \\ x^n(2) & P_{X^n}(x^n(2)) \\ \vdots & \vdots \\ x^n((a+1)^n) & \end{array} \right.$$

$x^n(1)$ 1 0

$x^n(2)$ 0 1

$x^n(3)$ 0 0

$x^n(4)$ 1 1

$\vdots$

0
1
00
01
10
11
000
001
010
$\vdots$
111
0000
$\vdots$

$\emptyset$
0
1
00
01
10
11
000
$\vdots$

$$\mathbb{E}\, k(X^n) = \sum_{i=1}^{(a+1)^n} k(x^n(i))\, P_{X^n}(x^n(i))$$

$x^n(i)$

$k(x^n(1)) = 0$

$k(x^n(2)) = 1$

$k(x^n(3)) = 1$

$4 = 2$

$5 = 2$

$6 = 2$

$7 = 2$

$8 = 3$

$\vdots$

$15 = 3$

$16 = 4$

$\vdots$

$31 = 4$

$\vdots$

$$k(x^n(i)) = \lfloor \log_2 i \rfloor$$

$$k(x^n(i)) = \lfloor \log_2 i \rfloor$$

$$p_{X^n}(x^n(1)) \geq p_{X^n}(x^n(2)) \geq \cdots$$

Claim : $p_{X^n}(x^n(i)) \leq \frac{1}{i}$

$p_{X^n}(1) \leq 1 \qquad p_{X^n}(2) \leq \frac{1}{2}$

$p_{X^n}(3) \leq \frac{1}{3}$

Assume: $p_{X^n}(2) > \frac{1}{2} \Rightarrow p_{X^n}(1) \geq p_{X^n}(2) > \frac{1}{2}$

$\Rightarrow p_{X^n}(1) + p_{X^n}(2) > \frac{1}{2} + \frac{1}{2} = 1$

EXPECTED LENGTH

$$\mathbb{E}\, k(X^n) = \sum_{i=1}^{(a+1)^n} k(x^n(i))\, p_{X^n}(x^n(i))$$

$$= \sum_{i=1}^{(a+1)^n} p_{X^n}(x^n(i)) \lfloor \log_2 i \rfloor$$

$$\leq \sum_{i=1}^{(a+1)^n} p_{X^n}(x^n(i)) \left\lfloor \log_2 \frac{1}{p_{X^n}(x^n(i))} \right\rfloor$$

$$\leq \sum_{i=1}^{(a+1)^n} p_{X^n}(x^n(i)) \log_2 \frac{1}{p_{X^n}(x^n(i))}$$

$$= \underbrace{\sum_{x^n} p_{X^n}(x^n) \log \frac{1}{p_{X^n}(x^n)}}_{H(X^n)}$$

ENTROPY OF p_{X^n}

In general: if $p_{X^n}(i) > \frac{1}{i}$ for some i

$$\Rightarrow \quad p_{X^n}(1) \geq p_{X^n}(2) \geq \cdots \geq p_{X^n}(i) > \frac{1}{i}$$

$$\sum_{j=1}^{i} p_{X^n}(j) = p_{X^n}(1) + p_{X^n}(2) + \cdots + p_{X^n}(i)$$
$$> \frac{1}{i} + \frac{1}{i} + \frac{1}{i} + \cdots + \frac{1}{i}$$
$$\geq i \times p_{X^n}(i) > i \times \frac{1}{i} = 1 \Rightarrow$$ NOT POSSIBLE

Contradiction!

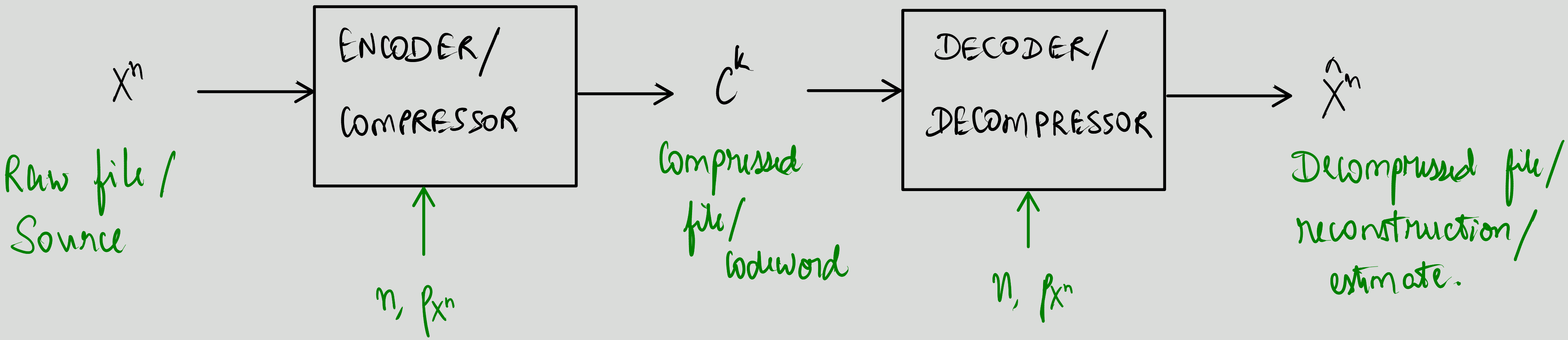
$$p_{X^n}(i) \leq \frac{1}{i} \Rightarrow i \leq \frac{1}{p_{X^n}(i)} = \frac{1}{p_{X^n}(x^n(i))}$$

$$R_{avg} = \lim_{n\to\infty} \frac{\mathbb{E}\, k(X^n)}{n} = \lim_{n\to\infty} \frac{H(X^n)}{n} \rightarrow \underline{\text{ENTROPY RATE}}$$

Can construct compressor for which R_{avg} = ENTROPY RATE.

For iid X^n,

$$\lim_{n\to\infty} \frac{H(X^n)}{n} = H(X) = \sum_{x\in\mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}$$



X^n
Raw file / Source
ENCODER/ COMPRESSOR
n, p_{X^n}
C^k
Compressed file/ codeword
DECODER/ DECOMPRESSOR
n, p_{X^n}
$\hat{X}^n$
Decompressed file/ reconstruction/ estimate.

Lossless source coding theorem:

① $X^n \sim iid(p_X)$

for every $\varepsilon > 0$, (fixed length compressor)

fixed length compression

Achievability: We can construct (ENC, DEC) st

$$R = H(X) + \varepsilon$$

$$P_e \to 0 \text{ as } n \to \infty$$

Converse: for every (ENC, DEC) with $R < H(X)$

$$P_e \to 1 \text{ as } n \to \infty$$

$$X^n \sim \text{iid}(p_X) \quad \Rightarrow \quad H(X^n) = n H(X) \Rightarrow \frac{H(X^n)}{n} = H(X)$$

$$H(X^n) = \sum_{x^n} p_{X^n}(x^n) \log_2 \frac{1}{p_{X^n}(x^n)}$$

$$= \sum_{x^n} \left(\prod_{i=1}^{n} p_X(x_i) \right) \log_2 \frac{1}{\left(\prod_{i=1}^{n} p_X(x_i) \right)} \qquad \text{(iid)}$$

$$= \sum_{x^n} \prod_{i=1}^{n} p_X(x_i) \sum_{j=1}^{n} \log \frac{1}{p_X(x_j)}$$

$$= \sum_{j=1}^{n} \sum_{x^n} \left(\prod_{i=1}^{n} p_X(x_i) \right) \left(\log \frac{1}{p_X(x_j)} \right)$$

$$= \sum_{j=1}^{n} \sum_{x_j} p_X(x_j) \log_2 \frac{1}{p_X(x_j)} = n \sum_{x} p_X(x) \log_2 \frac{1}{p_X(x)}$$

$H(X^n) = H(X_1) + H(X_2) + \cdots H(X_n)$ if X^n is iid

Entropy is non-negative

$X \sim p_X$

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{1}{p_X(x)}$$

By convention,

$p_X(x) = 0$

$$p_X(x) \log_2 \frac{1}{p_X(x)} = 0$$

$$\lim_{\substack{p \to 0 \\ p \downarrow 0}} p \log_2 \frac{1}{p} = 0$$

$$0 < p_X(x) \leq 1$$

$$\Rightarrow \quad 1 \leq \frac{1}{p_X(x)}$$

$$H(X) = \sum_{x} p_X(x) \log_2 \frac{1}{p_X(x)} \geq 0$$

Entropy: $H(X)$ $\downarrow$ actually a function of p_X & not X.

$H(p_X) \longrightarrow$ Better notation for entropy

Entropy is invariant to relabeling

$\mathcal{X} = \{0,1,2\}$. $p_X(0) = \frac{1}{2}$

$p_X(1) = \frac{1}{4}$

$p_X(2) = \frac{1}{4}$

$H(X) = 1.5$ bits

Entropy is a function only of the probability multiset $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$

$q_X(0) = \frac{1}{4}$

$q_X(1) = \frac{1}{2}$

$q_X(2) = \frac{1}{4}$

$H(q_X) = 1.5$ bits

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}$$ <u>bits</u>

$(\ ) \log_e (\ )$ <u>nats</u>

① $X \sim \mathrm{Unif}(\mathcal{X})$ $\quad |\mathcal{X}| = m$

$p_X(x) = \frac{1}{m} \quad \forall x \in \mathcal{X}$

$$H(X) = \sum_x \frac{1}{m} \log_2 \frac{1}{(1/m)}$$

$$= \sum_x \frac{1}{m} \log_2 m$$

$$= \log_2 m$$

$$H(X) = \log_2 |\mathcal{X}|$$

② $\mathcal{X} = \{1, 2, 3, \ldots\}$ $\quad \mathcal{Y} = \mathbb{Z}_{>0}$

Geometric rv: $\quad p_X(x) = \left(\frac{1}{2}\right)^x \qquad \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = 1$

$$H(X) = \sum_{x=1}^{\infty} \frac{1}{2^x} \log_2 \frac{1}{1/2^x} = \sum_{x=1}^{\infty} \frac{x}{2^x} = 2 \text{ bits.}$$

$$\sum_{x=1}^{\infty} \frac{x}{2^x} \qquad\qquad \sum_{n=1}^{\infty} p^n = \frac{p}{1-p}$$

$$\sum_{n=1}^{\infty} np^n = p\sum_{n=1}^{\infty} np^{n-1} = p\,\frac{d}{dp}\left(\sum_{n=1}^{\infty} p^n\right)$$

$$= p\,\frac{d}{dp}\left(\frac{p}{1-p}\right) = p\left(\frac{1}{1-p} + -\frac{p}{(1-p)^2}(-1)\right)$$

$$= p \times \frac{1-p+p}{(1-p)^2}$$

$$= \frac{p}{(1-p)^2}$$

$X_i \in \mathcal{X}$ $\qquad$ $\mathbb{E}\,k(X^n) \approx H(X^n)$ $\qquad$ $|\mathcal{X}| = m$

X^n compressed $\underline{\underline{C^k}}$ $\underline{\underline{\mathbb{E}\,k}} \approx H(X^n) = nH(X)$ (iid) $\leq n\log|\mathcal{X}|$ $\qquad$ $\underline{\underline{n\lceil \log_2 m\rceil}}$

iid Geometric$(1/2)$ seq $\rightarrow$ $\underline{\underline{2n}}$ bits on avg

1	000
2	001
3	010
4	011
5	100

$$\lceil \log_2 n \rceil$$

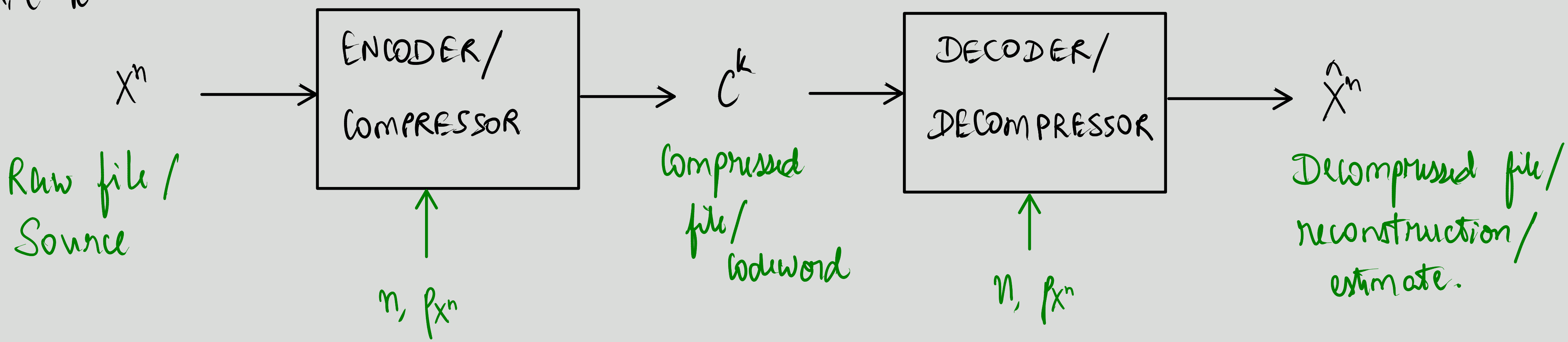
Fixed-length compression

$k \rightarrow$ fn of n, p_{X^n}

$X_1 \cdots X_n$

$X_i \in \mathcal{X}$

$c^k \in \{0,1\}^k$



$$P_e = \Pr[\hat{X}^n \neq X^n]$$

k – fixed depends only on n, p_X

$X^n \sim$ iid(p_X) $\qquad X_i \sim p_X$

$R = \dfrac{k}{n}$, P_e small.

$n \gg 1$ $\qquad n \rightarrow \infty$ $\qquad \lim_{n \to \infty} \dfrac{k}{n} = R$ small, $P_e \rightarrow 0$ as $n \rightarrow \infty$

Compressing Bernoulli(p) sequences

$$p_X(1) = p \quad , \quad p_X(0) = 1-p .$$

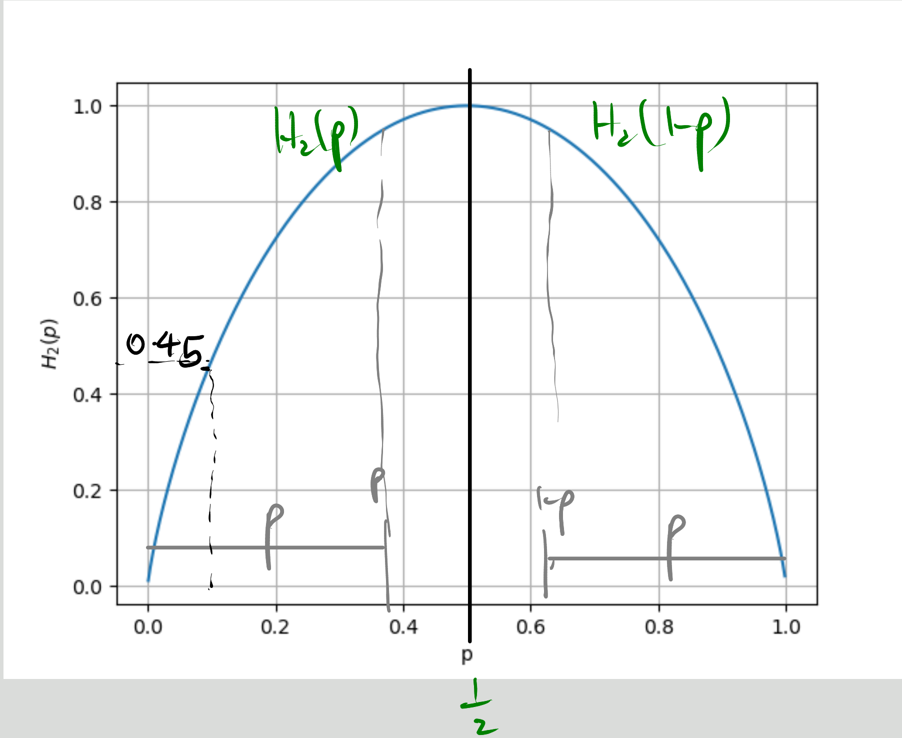
$$X^n = 0\,1\,0\,1\,1\,0\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0$$

$$\downarrow$$

$$C^k \qquad 0\,0\,0\,0\,1\,0\,0\,1\,0$$

$$R_{opt} = H(X)$$

$$= p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$



$$R_{opt}(p) = H_2(p) \longrightarrow \text{Binary entropy } (p)$$

as $p \to 0$ $\quad p \log_2 \frac{1}{p} \to 0 \quad , \quad (1-p) \log_2 \frac{1}{1-p} \longrightarrow 0 \quad \Rightarrow \quad H_2(p) \to 0$

$p \to 1 \qquad\qquad H_2(p) \to 0$

GOAL: Construct ENC, DEC st

① $\frac{k}{n} \to H(X)$ as $n \to \infty$

② $P_e = \Pr[\hat{X}^n \neq X^n] \to 0$ as $n \to \infty$.

Structure of ANY fixed-length compressor

$X^n \sim$ iid $\mathrm{Ber}(0.1)$ $\qquad H_2(0.1) = 0.45$

$n \to 0.45n$

10^7

$X^n \longrightarrow C^{0.45n}$ $\qquad 10^{-6}$

$\underline{X^{1000}} \to C^{450} \to \underline{\hat{X}^{1000}}$ $\qquad \Pr[\hat{X}^{1000} \neq X^{1000}] \leq 10^{-2}$

EXTREME: $p=0$ $\quad n$ $\qquad 0000 \cdots 0$ $\quad k=0$

$\longleftarrow n \longrightarrow$

$p=1$ $\qquad 11\cdots 1$ $\quad k=0$

X^n: ENC, DEC, know # 1's $= 1$, n

$n = 100{,}000$

ENC: location of the 1.

$k = \lceil \log_2 100{,}000 \rceil = \lceil \log_2 n \rceil \ll n$

ASSUME: $\Pr[\#\text{ 1's in } X^n > 1] \le 1/n$

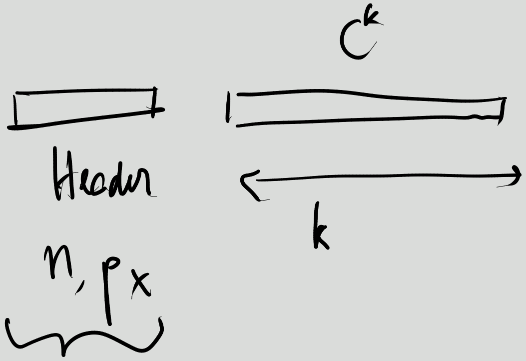
$\frac{k}{n} = \frac{\lceil \log_2 n \rceil}{n} \xrightarrow[n\to\infty]{} 0$

$P_e = \le \frac{1}{n} \longrightarrow 0$ as $n \to \infty$

$p = 0 \qquad k = 0 \qquad p_c = 0$

Assumptions: (n, p) known

Compressed file:



$O(\log n)$

Can be ignored if $n \gg 1$.

$a \to 0$

$b \to 10$

$c \to 110$

$d \to 1100$

$e \to 1110$

$\lceil \log_2 5 \rceil \simeq 3$ bits

2^8

$X^n \sim$ iid Ber(p)

$\#$ 1's $\approx np$

$\Pr[\,|\#1\text{'s in } X^n - np| > n\varepsilon] \longrightarrow 0$ as $n \to \infty$

$\Pr[\,|(\text{fraction of 1's in } X^n) - p| > \varepsilon] \longrightarrow 0$ as $n \to \infty$

(WLLN)

$2^n - 2^{nH_2(p)}$

$S_{\varepsilon,p,n} = \{ x^n \in \{0,1\}^n : np(1-\varepsilon) \le \mu_1(x^n) \le np(1+\varepsilon) \}$

$|S| \approx 2^{nH_2(p)}$

$\ll 2^n$

$\Pr[X^n \in S] \longrightarrow 1$ as $n \to \infty$

$x^n(1)$ 000...0

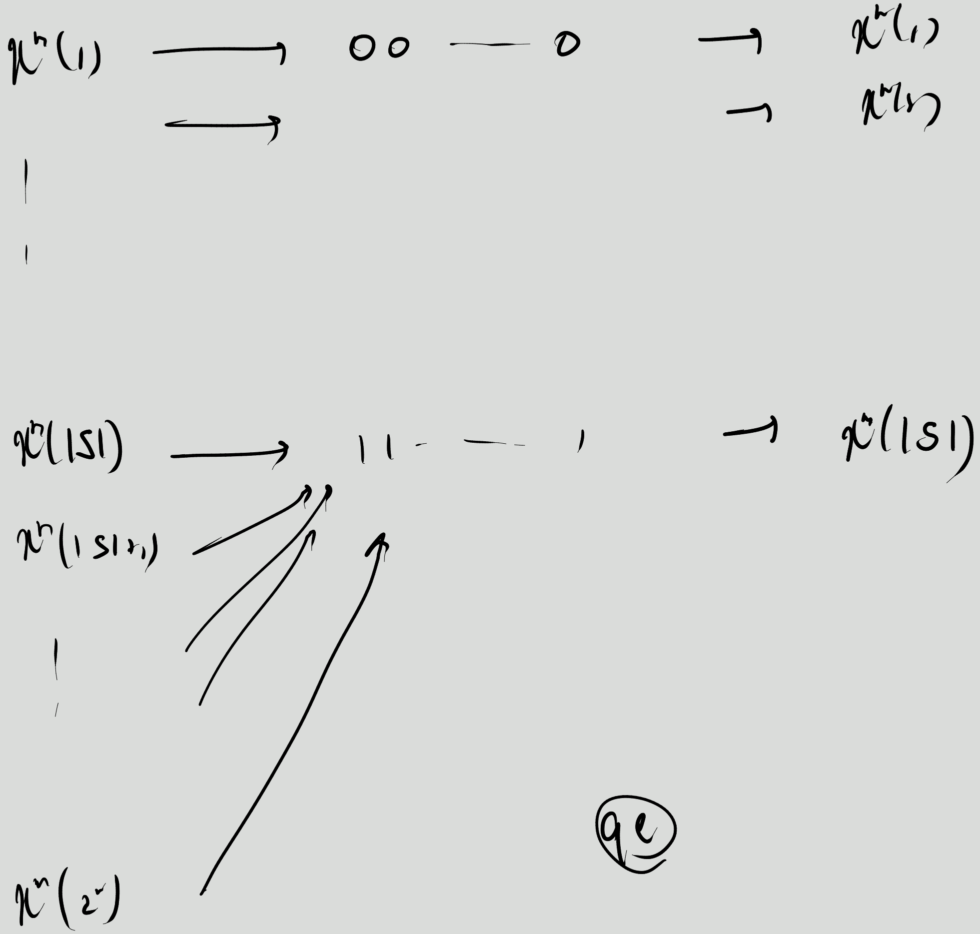
$\vdots$ 00...1

$\vdots$

$x^n(|S|)$ 11...1

$k = \lceil \log_2 |S| \rceil$

$\approx nH(X) = \underline{\underline{nH_2(p)}}$



qe

an

as

q2