EE 5390

Source Coding

Logistics

Google classroom: Invites shared

Course webpage

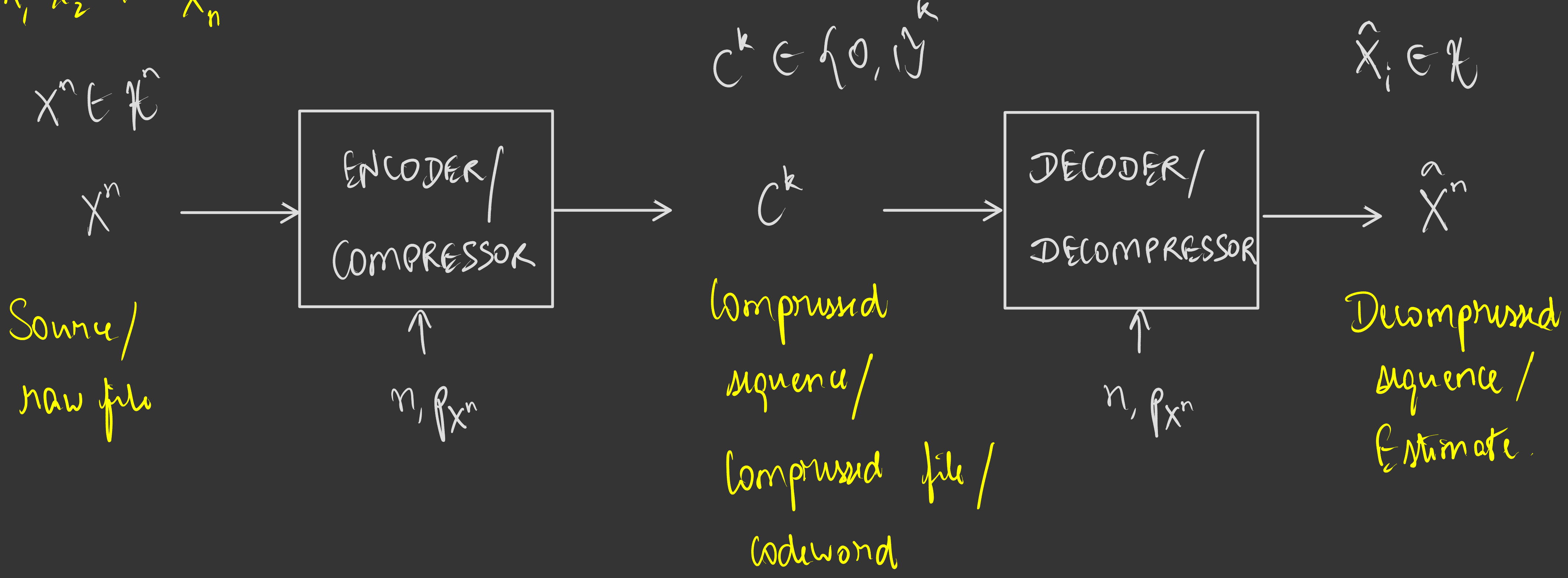
This course

Data compression: fundamental limits and algorithms

Probabilistic framework

General setup

$X_1, X_2 \dots X_n$ $X_i \in \mathcal{X} \rightarrow$ Alphabet



$X^n \in \mathcal{X}^n$

$X^n \rightarrow$ ENCODER/ COMPRESSOR $\rightarrow C^k \rightarrow$ DECODER/ DECOMPRESSOR $\rightarrow \hat{X}^n$

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

$\hat{X}_i \in \mathcal{X}$

n, P_{X^n} n, P_{X^n}

Source/ raw file

Compressed sequence/ Compressed file/ codeword

Decompressed sequence/ Estimate.

Text file: ASCII characters $\mathcal{X} = \{$ ASCII chars $\}$

$|\mathcal{X}| = 2^8$

English text: Words/ Sentences $\mathcal{X} = \{$ all English words, punctuation marks $\}$

English text : Structure

$\downarrow$

random process

Source : $X^n \sim p_{X^n} \rightarrow \underline{\underline{\text{Given}}}$

<u>Model</u> : X^n

① n is known

② p_{X^n} is known

Classification of compression algorithms

ENC: $X^n \mapsto C^k$ DEC: $C^k \mapsto \hat{X}^n$

1. Based on fidelity

① Lossless

② Lossy

① Lossless

① Zero-error: $\hat{X}^n = X^n$

text compression zip, rar, gzip, 7zip

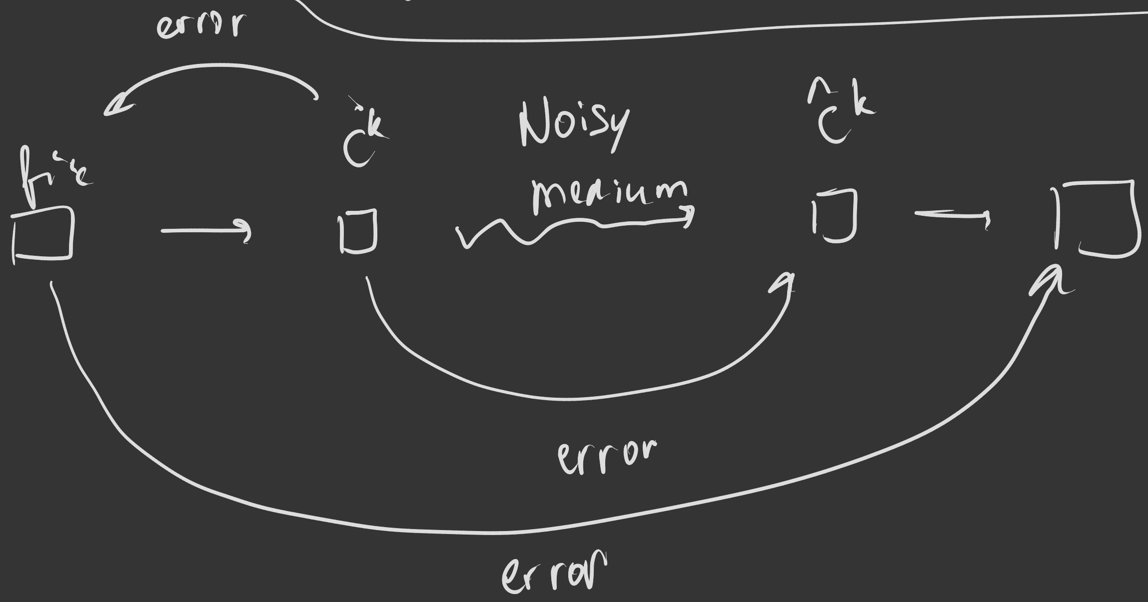
② Almost lossless/ Vanishing error probability

probability of error $P_e = Pr\left[\hat{X}^n \neq X^n\right]$ small

$P_e \to 0$ as $n \to \infty$

k

$P_e < \varepsilon$



$P_e^{(1)} = Pr\left[\hat{C}^k \neq C^k\right]$

$P_e^{(2)} = Pr\left[DEC(C^k) \neq X^n\right]$

$P_e^{overall} \leq P_e^{(1)} + P_e^{(2)}$

② Lossy

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

jpg , mp3 , mp4

jpg few MB → Lossy

upload low res image kB

$\hat{X}^n \neq X^n$ $d(\hat{X}^n, X^n) \le \delta$

$\uparrow$

Distortion measure

Tradeoff b/w compressed filesize & δ

$\delta \uparrow \Rightarrow$ resolution smaller distortion:

$$d(\hat{X}^n, X^n) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2\right] \leq \delta$$

Mean Squared error

MSE

2. Based on compressed length

① Fixed length

k is fixed for a given (n, p_{X^n})

② Variable length

k can depend on n, p_{X^n}, X^n

Zip, rar, 7zip, gz

Compression algorithms

Fidelity/Quality

① Lossless

$\underline{\hat{X}^n} \approx \underline{X^n}$

② Lossy

$d(\hat{x}^n, x^n) \leq \delta$

Length

① fixed length

k - fn of p_{X^n}, n

② Variable length

k - fn of n, p_{X^n}, X^n

Example: Fixed-length compression

	x^n	$p(x^n)$	$ENC(x^n)$
①	000	$1/2$	00
②	001	$1/4$	01
	010	$1/8$	10
	011	$1/16$	11
	100	$1/32$	00
	101	$1/64$	01
	110	$1/128$ $\quad 1/2^i$	10
	111	$1/128$	11

Source

	c^k	$DEC(c^k)$
000 100	00	000
001 101	01	001
010 110	10	010
	11	011

$$R = \frac{k}{n} = \frac{2}{3}$$

$$P_e = P(100) + P(101) + P(110) + P(111)$$

$$= \frac{1}{16}$$

$n = 3 \qquad k = 2$

$$ENC(x_1 x_2 x_3) = x_2 x_3 \qquad DEC(y_1 y_2) = 0 y_1 y_2$$

Example: Variable-length compression

	x^n	$p(x^n)$	$ENC(x^n)$	
①	000	1/2	0	000
②	001	1/4	10	001
③	010	1/8	110	010
	011	1/16	1110	⋮
	100	1/32	11110	
	101	1/64	111110	
	110	1/128	1111110	
	111	1/128	1111111	

$k_{max} = 7$ bits.

$P_e = 0$

ENC: for ith seq: $(i-1)$ 1's 0

$$\mathbb{E}k = \sum_{x^n} p(x^n)\, k(x^n) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \cdots = 1.98 \text{ bits}$$

Evaluating the performance of a compression scheme

① Fixed-length compressor

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

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

$\left.\begin{array}{l} R \downarrow \\ \\ P_e \downarrow \end{array}\right\} \; n \gg 1$

$\lim_{n\to\infty} R$ as small as possible

$\lim_{n\to\infty} P_e = 0$

$P_e = \frac{1}{n}$

② Variable-length compressor : $P_e = 0$

Expected length $= \mathbb{E}k$

Average rate $= \frac{\mathbb{E}k}{n}$

$R_{avg} \downarrow$

as small as possible.

The lossless source coding theorem for memoryless sources (fixed-length compression)

(Claude Shannon)

If $X^n \sim \text{iid}(p_X)$, then there exist (ENC, DEC) st

① $\lim\limits_{n\to\infty} \frac{k}{n} = H(X) + \varepsilon$ $\quad = \sum\limits_{x\in\mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}$

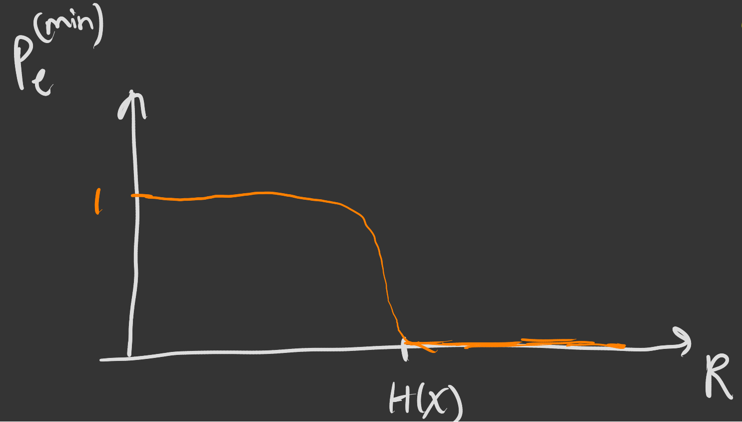
0.1

② $\lim\limits_{n\to\infty} P_e = 0$

For any (ENC, DEC) satisfying $\lim\limits_{n\to\infty} \frac{k}{n} < H(X)$,

$\lim\limits_{n\to\infty} P_e = 1$

Fix p_X



If $X^n \sim P_{X^n}$ stationary & ergodic

$$R_{min} = \lim_{n \to \infty} \frac{H(X^n)}{n} \rightarrow \text{Entropy rate}$$

$X^n \sim iid(P_X)$

$$\mu_a(X^n) = \text{freq of occurrence of } a \text{ in } X^n$$

$$\approx P_X(a)$$

$X^n \rightarrow \quad X_1 \sim P_X \quad Ber(1/2) \qquad X_2 = X_3 = \cdots\cdots = X_1$

$\forall i, \quad P_{X_i}(a) = P_{X_1}(a) = \underline{\underline{\frac{1}{2}}}$

$\underline{\mu_a(X^n) = 0}$ or 1

NOT ERGODIC

$$Pr[\mu_a(X^n) \approx P_{X^n}(a)] \ll 1$$

$$R_{min} = \lim_{n\to\infty} \frac{H(X_1 \cdots X_n)}{n}$$

$X^n \sim iid(p_X)$

$$= \lim_{n\to\infty} \frac{\cancel{n} H(X)}{\cancel{n}}$$

$$= H(X) = H(X_1)$$

for iid X^n,

$$H(X_1 \cdots X_n)$$

$$= \sum_{i=1}^{n} \underbrace{H(X_i | X_1 \cdots X_{i-1})}_{H(X_i) = H(X_1) = H(X)}$$

$$=$$

$$= n H(X)$$

The lossy source coding theorem for memoryless sources

Suppose we want $\mathbb{E}d(X^n, \hat{X}^n) = \sum_{i=1}^{n} \mathbb{E}d(X_i, \hat{X}_i) \le n\delta$

$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ Distortion fn

If $X^n \sim iid(p_X)$, then there exist (ENC, DEC) st

Rate-distortion

① $$\lim_{n\to\infty} \frac{k}{n} = R(\delta) = \boxed{\min_{\substack{p_{\hat{X}|X}: \\ \mathbb{E}d(X,\hat{X}) \le \delta}} I(X;\hat{X})} = f(\delta, d, p_X)$$

② $$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}d(X^n, \hat{X}^n) = \delta$$

For any (ENC, DEC) satisfying $\lim_{n\to\infty} \frac{k}{n} < R(\delta)$

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}d(X^n, \hat{X}^n) > \delta$$

Preliminaries

① Definitions of $H(X)$, $H(X,Y)$, $H(Y|X)$, $I(X;Y)$, $I(X;YZ)$, $I(X;Y|Z)$, $D(p\|q)$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

② Chain rules

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

$$I(X;YZ) = I(X;Y) + I(X;Z|Y)$$

③ $D(p\|q) \geq 0$

$H(X) \geq 0$

$I(X;Y) \geq 0$