7 pages

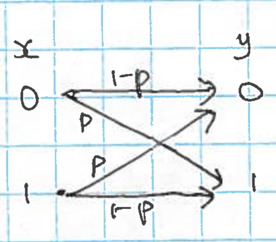
Def. A classical channel N is specified by:

- an input alphabet X
- an output alphabet Y
- a distribution $p(y|x)$ for each $x \in X$.

eg 1 Binary symmetric channel (BSC)

$X = Y = \{0,1\}$, input "flipped wp p"

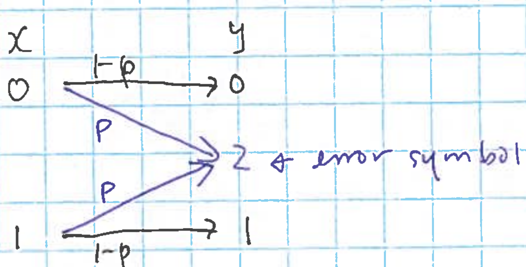
$p(y=i|x=i) = 1-p$.
$p(y \neq i|x=i) = p$.



eg 2 Erasure channel.

$X = \{0,1\}$, $Y = \{0,1,2\}$.

Input sent wp $1-p$
replaced by "2" wp p



Main focus:

① Asymptotic rate of comm
- can use channel many (n) times
- allow a very small error rate

② Discrete memoryless channels DMCs.
ie each use independent & identical:

For input $= x_1 \; x_2 \; \cdots \; x_n$
output $= y_1 \; y_2 \quad y_n$ wp $\prod_{i=1}^{n} p(y_i | x_i)$.

Aside: sensible channels not of the above type.

eg 1. Missing-symbol-channel
$x_1 x_2 \cdots x_n \longrightarrow \underbrace{y_1 y_2 \cdots y_m}_{\text{obtained by } x_1 \cdots x_n \text{ by deleting } n-m \text{ symbols.}}$, [illegible]

Don't know which symbols are missing.

eg 2. channel.
$x_1 x_2 \cdots x_i x_{i+1} \cdots x_n \rightarrow x_1 x_2 \cdots x_{i+1} x_i \cdots x_n$

Symbols can emerge out of order

eg 3. Burst errors.
$x_1 x_2 \cdots x_n \rightarrow x_1 x_2$ ~~$\cdots$~~ x_n.
missing a large contiguous block of symbols.
eg someone pulled off a page in a book

DMC from now on:

Use error correcting codes:

eg. Repetition code

message		codeword	
0	$\longrightarrow$	$00\cdots0$	} majority decoding
1	$\longrightarrow$	$11\cdots1$	
		$\leftarrow k$ times $\rightarrow$	

"The code": set of legitimate code words (a subset of all possible inputs).

Rate $= \frac{1}{k}$

Prob error $\approx p^k$ for erasure channel

eg. Hamming code $\forall r \geq 2$

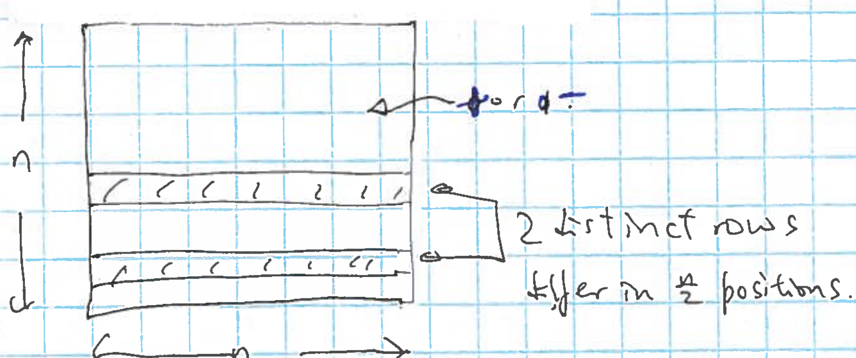
Encodes $K = 2^r - r - 1$ in $n = 2^r - 1$ bits

$\text{Rate} = \frac{k}{n} = 1 - \frac{r}{2^r - 1}$

Can correct 1 error in BSC.

eg. Hadamard code.

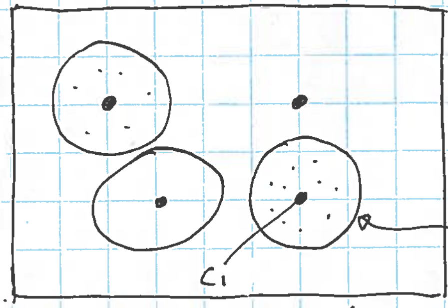
Take Hadamard matrix:



Bit flip errors (up to $\frac{n}{4} - 1$) can be corrected. Rate $\frac{k = \log n}{n} \to 0$ as $n \to \infty$.

Geometric interpretation.

For simplicity, take $X = Y$.



• ← code words

: Hamming spheres of radius k

all output strings within k errors from c_i

$X^{\otimes n} = Y^{\otimes n}$

Using only codewords to represent messages, they can be recovered if at most k errors happen and if Hamming spheres don't overlap.

If few enough codewords

Qn:

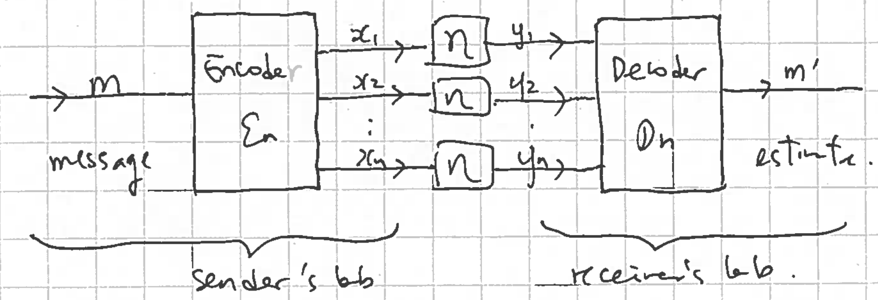
① for a fixed message size, to have smaller & smaller error prob, need bigger & bigger codes

but that means more & more errors.

Can error prob → 0?
With the rate → 0?

② for growing message size, even if prob for each segment → 0, can the entire message be sent correctly up → 1?

Sending messages through n uses of a noisy channel:



An "(M, n)" code consists of

① an index set $\mathcal{M} = \{1, 2, \ldots, M\}$

② an encoding function $\mathcal{E}_n : \mathcal{M} \to X^{\otimes n}$

③ a decoding function $\mathcal{D}_n : Y^{\otimes n} \to \mathcal{M}$.

} Will see why randomness not needed.

- "The code" $= \{ \underbrace{\mathcal{E}_n(1), \mathcal{E}_n(2)}_{\text{codewords}}, \ldots, \mathcal{E}_n(M) \}$

- Rate of (M, n) code $= \frac{1}{n} \log M$.

  Given (M, n) code

- for each m, m' follows a distribution. (Randomness introduced by the noisy channel(.))

- Define $\underset{\text{wrt code.}}{P_e}(m) := \text{Prob}(m' \neq m \mid m)$.

  $\underset{\text{wrt code.}}{P_e} := \max_{m \in \mathcal{M}} P_e(m)$ worse case error

  $\mathbb{E}\, P_e := \frac{1}{M} \sum_{m \in \mathcal{M}} P_e(m)$ average error.

Def [achievable rate]:

For a channel $\mathcal{N}$, a rate R is achievable if $\exists$ sequence of $(\underbrace{\lfloor 2^{nR} \rfloor}_{M}, n)$ codes C_n s.t. $P_e(C_n) \to 0$ as $n \to \infty$

Def The capacity of $\mathcal{N}$, $C(\mathcal{N})$, is the supremum over achievable rates.

← best # bit transmitted per channel use

NB: If $C(\mathcal{N}) > 0$, the entire message (longer & longer, $\approx nR$ bits) comes out correctly (in each symbol) almost surely!

Thm (Shannon's noisy coding theorem)

$$C(\mathcal{N}) = \max_{p(x)} I(X:Y)$$

where $p(xy) = p(x) \cdot \underbrace{p(y|x)}_{\text{where the channel comes in.}}$

NB1. LHS is an asymptotic quantity ↗ no direct relation.
RHS is a "single-letter" formula — involving only 1 copy of $\mathcal{N}$.

NB2. Since works in worst case, message NEED not follow any distribution. (will see where p(x) appears.)

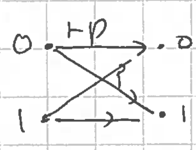
eg. 1. $C(N) = 0$ iff $p(xy) = p(x) \cdot p(y|x) = p(x) \cdot p(y)$.

iff $\underbrace{p(y|x) = p(y)}$, X, Y indep.

Channel discarding input, & draws y according to $p(y)$.

So all other channels have $C(N) > 0$!

eg 2. BSC. N_p



$$I(X:Y) = H(Y) - \underbrace{H(Y|X)}_{= H(p)}$$

$H(Y|X=x)$ indep of x

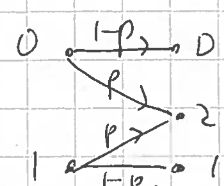
$\therefore H(Y|X)$ - - - $p(x)$.

$\therefore$ choose $p(x)$ to max $H(Y)$

optimal $p(x)$: $p(0) = p(1) = \frac{1}{2}$, $H(Y) = 1$

$\therefore C(N_p) = 1 - H(p)$.

eg 2 Erasure channel $\mathcal{E}_p$



$$I(X:Y) = H(X) - \underbrace{H(X|Y)} = 1 - p.$$

$H(X|0) = H(X|1) = 0$

$H(X|2) = 1$.

$\therefore H(X|Y) = p$.

$\therefore$ choose $p(x)$ to max $H(X)$, $p(0) = p(1) = \frac{1}{2}$

Note diff way to explain $I(X:Y)$.

Note zero error channel capacity.

How to prove (*) ?

① Prove a converse :

ie at any higher rate, $\overline{P_e} \not\to 0$

[Thus $C(N) \le \max_{p(x)} I(X:Y)$]

② Prove a direct coding Theorem :

ie given any $p(x)$,

Show $\exists$ codes achieving the rate $I(X:Y)$

+ error analysis $\overline{P_e} \to 0$

[Thus $C(N) \ge \max_{p(x)} I(X:Y)$]

holds for completely different reasons.

(7-8 pages)

Direct coding theorem:

- Need to show $\exists$ (M, n) codes C_n with:
  - rate $\frac{1}{n} \log M \geq I(X;Y) - \delta_n$ $\quad$ $\delta_n \to 0$
  - error $P_e(C_n) \to 0$

- Strangely, we need not find these codes.

- Ideas: for each n, a code C_n is generated via a random process

  Show $\underset{C_n}{E}\ E P_e(C_n) \to 0$ while $\delta_n \to 0$

  over messages

  So, some $\tilde{C}_n$ has $E P_e(C_n) \to 0$.
  Derive a modified $\hat{C}_n$ with $P_e(\hat{C}_n) \to 0$.

Given any n, M, $p(x)$, the random process generating C_n is:

for $i = 1 \ldots M$
$\quad j = 1 \ldots n$

pick x_{ij} iid $\sim p(x)$.

Let
$$\left.\begin{aligned} C_1 &= x_{11}\, x_{12} \ldots x_{1n} \\ C_2 &= x_{21}\, x_{22} \quad\; x_{2n} \\ &\vdots \\ C_M &= x_{M1}\, x_{M2} \ldots x_{Mn} \end{aligned}\right\}$$ the M codewords defining the code C_n

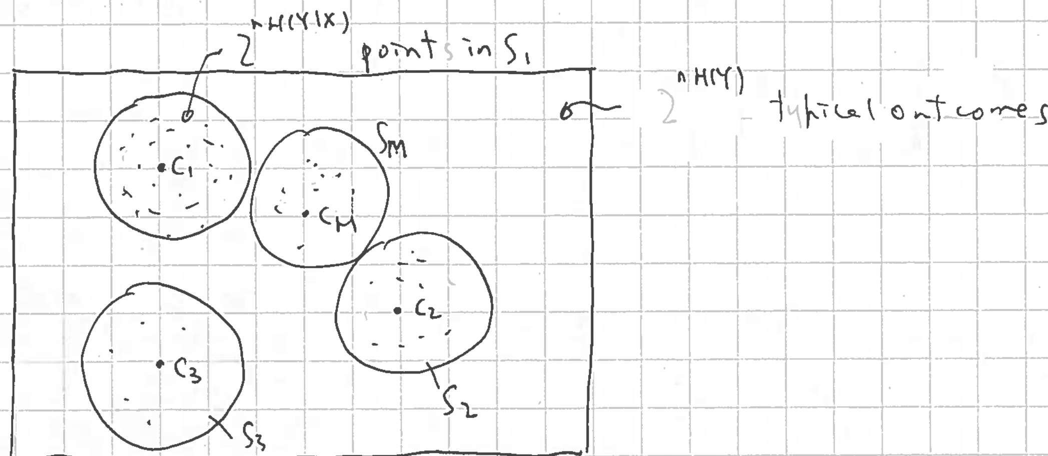
Why $P_e(C^n) \to 0$ if $M < 2^{n(I(X:Y)-\delta_n)}$?

(take $\Omega_X = \Omega_Y$ for simplicity).

- Output Y^n is iid $\sim p(y)$
- There are $2^{nH(Y)}$ typical outcomes.

If message is i, C_i transmitted via $N^{\otimes n}$

- then there are $\approx 2^{nH(Y|X)}$ outcomes, centred at C_i, forming a set S_i



- If $2^{nH(X|Y)} \cdot M \ll 2^{nH(Y)}$

  then $S_1, S_2 \dots S_M$ don't overlap much.

- Bob can find out which C_i was sent by Alice

  [by checking which S_i his output y^n belongs to]

Actual analysis requires the following:

Def: [Jointly typical sequence].

Given dist^n $p(xy)$, drawn iid n times,

$x^n y^n$ is ε-jointly typical if

(a) $\left|\frac{-1}{n} \log p(x^n) - H(X)\right| \leq \varepsilon$ (x^n typical)

(b) $\left|\frac{-1}{n} \log p(y^n) - H(Y)\right| \leq \varepsilon$ (y^n typical)

(c) $\left|-\frac{1}{n} \log p(x^n y^n) - H(XY)\right| \leq \varepsilon$ ($(xy)^n$ typical)

Def [Jointly typical set)

$A_{n,\varepsilon} = \{x^n y^n \;\; \varepsilon\text{-jointly typical}\}$

Obs: Within $A_{n,\varepsilon}$, prob$(y^n|x^n)$ etc all tightly concentrated.

Thm [Joint AEP].

Using above defs, $\forall \delta > 0 \;\; \forall \varepsilon > 0 \;\; \exists n_0$ s.t. $\forall n \geq n_0$

① $\Pr(A_{n,\varepsilon}) \geq 1-\delta$

② $(1-\delta)\, 2^{n(H(XY)-\varepsilon)} \leq |A_{n,\varepsilon}| \leq 2^{n[H(XY)+\varepsilon]}$

③ If we draw $(x^n)(y^n)$ according to some other dist^n

$q(x^n y^n) = p(x^n) \cdot p(y^n)$

then $\Pr_q (x^n y^n \in A_{n,\varepsilon}) \leq 2^{-n(I(X;Y) - 3\varepsilon)}$

$2^{-n(I(X;Y)+3\varepsilon)} \leq$

Pf = ex.

More specifically:

① A random code C_n is generated.

② C_n is told to Alice & Bob ← They stick with this given code. They know what $\mathcal{N}$ is.

③ A message i is drawn uniformly from $\{1, 2, \ldots, M\}$.

④ Alice sends C_i through $N^{\otimes n}$

⑤ Bob receives $Y^n \sim \Pr(y^n \mid C_i)$. $x_{i1} x_{i2} \ldots x_{in}$.

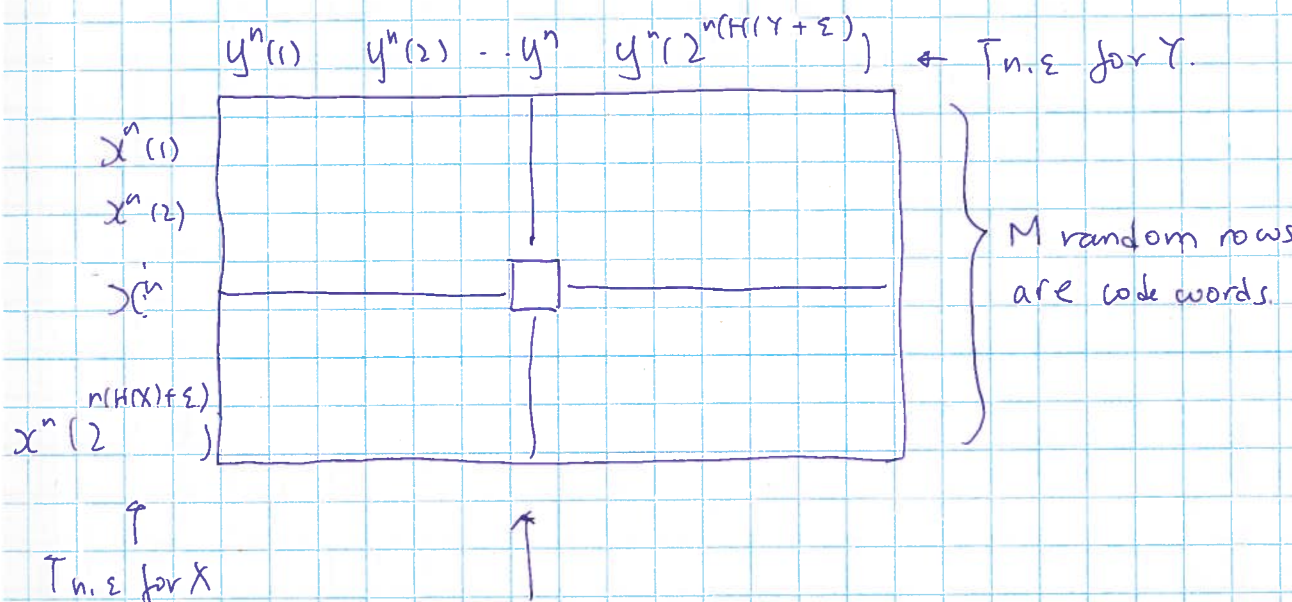
⑥ Bob decodes the outcome y^n:

If $\exists !\, j$ s.t. $C_j y^n \in A_{n,\varepsilon}$, Bob outputs j

Otherwise Bob outputs "failed".

"Joint typicality decoding"

If we make a table:



If $x^n y^n \in A_{n,\varepsilon}$, put a "1"

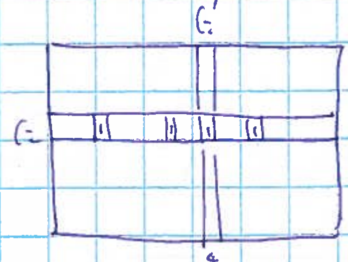
else - - - "0".

Roughly $2^{n\,H(X,Y)}$ 1's

When message is i, Alice transmits row "C_i"

By JAEP ① w.p. $\geq 1-\delta$, the following holds:

Bob receives C_i' s.t. $C_i C_i' \in A_{n,\varepsilon}$.



What Bob sees.

If column for C_i', restricted to the code, has a unique "1", that "1" is in the row C_i.

$\therefore$ Bob's output $j = i$.

Only error comes from:

① JAEP ① fails.

② There is some C_k, $k \neq j$ s.t. $C_k C_i' \in A_{n,\varepsilon}$.

But C_i' depends on C_i, indep of C_k.

$\therefore$ $\Pr(C_k C_i' \in A_{n,\varepsilon}) \leq 2^{-n(I(X;Y)-3\varepsilon)}$ by JAEP ③.

over choice of C_n.

$$\therefore \underset{C_n}{\mathbb{E}} \; E P_e(C_n)$$

$$= \underset{C_n}{\mathbb{E}} \frac{1}{M} \sum_i P_e(C_i)$$

$$= \frac{1}{M} \sum_i \underbrace{\underset{C_n}{\mathbb{E}} P_e(C_i)}$$

$$\leq \delta + \sum_k P_r(C_k C_i' \in A_{n,\varepsilon})$$

$$\leq \delta + M \cdot 2^{-n(I(X;Y)-3\varepsilon)} \qquad \leftarrow \text{Indep of } i$$

$$= \delta + M \cdot 2^{-n(I(X;Y)-3\varepsilon)}.$$

$\therefore$ if rate $R = \frac{1}{n} \log M < I(X;Y) - 3\varepsilon$

for large n, $M \cdot 2^{-n(I(X;Y)-3\varepsilon)} < \eta$

$\delta < \eta$

$$\underset{C_n}{\mathbb{E}} \; E P_e(C_n) < 2\eta$$

So $\exists \hat{C_n}$ s.t. $EP_e(C_n) < 2\eta$.

From $\tilde{C}_n$, get a code $\tilde{\tilde{C}}_n$ with $P_e(\tilde{\tilde{C}}_n) \leq 2 \cdot E P_e(\tilde{C}_n)$

Take those $\frac{M}{2}$ code words in $\tilde{C}_n$ with prob of error less than the median.

$\tilde{\tilde{C}}_n$ sends 1 fewer bit than $\tilde{C}_n$.

rate goes down by $\frac{1}{n} \to 0$.

Techniques:

- Random codes
- Symmetry arguments
- Existential proof
- Expunging worse codewords to bootstrap EP_e to P_e.

3 pages

Converse:

$$nR = H(M) = \underbrace{H(M|Y^n)}_{\text{① } \le 1 + P_e nR} + \underbrace{I(M:Y^n)}_{\text{② } \le I(\mathcal{E}_n(M):Y^n) \text{ ③ } \le n \max_{p(x)} I(X:Y)}.$$

"Fanos ineq": essentially if $P_e \to 0$, can deduce M from Y^n.

"Data processing ineq"

① Thm [Fanos ineq]:

Let $P_e = \text{prob}(X \neq Z)$, $Z = f(Y)$

Ω = sample space of X.

Then $H(P_e) + P_e(\log|\Omega| - 1) \geq H(X|Y)$.

binary entropy fn

Pf: Define new rv E s.t.

$E = 0$ if $X = Z$

$E = 1$ otherwise.

$$H(EX|Y) = H(X|Y) + \underbrace{H(E|XY)}_{\text{NB } 0}.$$

$H(EXY) - H(Y)$ $\quad$ $H(XY) - H(Y)$ $\quad$ $H(EXY) - H(XY)$.

exchange EX $\because$ sym on LHS.

$$= H(E|Y) + H(X|EY)$$

$$\therefore H(X|Y) = H(E|Y) + H(X|EY)$$

$$\leq H(E) + \sum_y p(y) \left[P_e H(X|E=1\ Y=y) + (1-P_e) H(X|E=0\ Y=y) \right]$$

$$\leq H(P_e) + P_e \log(|\Omega| - 1)$$

Replacing X by M

Y by Y^n

$|\Omega|$ by 2^{nR}

$H(M|Y^n) \leq H(P_e) + P_e \cdot nR$. gives ①

$\leq 1 + P_e nR$.

② Thm [Data Processing ineq (DPI)]:

If 3 r.v.'s E, F, G form a Markov Chain "$E \to F \to G$"

[ie $I(E:G|F) = 0$]

Then $I(E:F) \geq I(E:G)$.

Pf: $I(E:FG) = \underbrace{H(E)}_{} - H(E|FG)$

$= I(E:G) + \underbrace{H(E|G) - H(E|FG)}_{}$

$= I(E:G) + I(E:F|G)$

$\because$ LHS sym wrt exchanging F & G.

$\therefore$ RHS same, $I(E:G) + \underbrace{I(E:F|G)}_{\geq 0 \text{ by SSA}} = I(E:F) + \underbrace{I(E:G|F)}_{=0 \text{ for } E\to F\to G}$.

$\therefore I(E:G) \leq I(E:F)$

(3). Lemma: Let $Y^n = N^{\otimes n}(X^n)$

Then $I(X^n : Y^n) \leq \sum_{i=1}^{n} (X_i : Y_i)$.

NB neither of X^n nor Y^n need to be iid.

e.g. $X^n = 111100$

$p \quad 1-p \quad \dots \quad 1-p \quad p$

$Y^n = 0/1 \quad \dots \quad 0/1$

Pf: $I(X^n : Y^n) = H(Y^n) - H(Y^n | X^n)$

$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y_1 Y_2 \dots Y_{i-1} X^n)$ Chain rule

$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | X_i)$ Y_i only depends on X_i

$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i)$ SA

$= \sum_{i=1}^{n} I(X_i : Y_i)$.

eg. Back comm doesn't help.

elaborate.