Def. A classical channel $N$ is specified by:

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

**Example 1: Binary symmetric channel (BSC)**

$X = Y = \{0, 1\}$, input "flipped with probability $p$".

- $p(y = 0 | x = 1) = 1 - p$
- $p(y = 1 | x = 0) = p$.

**Example 2: Erasure channel**

$X = \{0, 1\}$, $Y = \{0, 1, \text{?}\}$.

Input sent up $1 - p$ replaced by "?" with probability $p$.

**Diagrams:**

1. Binary symmetric channel (BSC) diagram showing the probability of error.
2. Erasure channel diagram showing the probability of erasure and the correct symbol replacement.
Main focus:

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

2. Discrete memoryless channels (DMC):
   - each use independent & identical:

   \[
   \text{Input} = X_1, X_2, \ldots, X_n \quad \text{Output} = Y_1, Y_2, \ldots, Y_n \quad \text{Up to} \quad \frac{1}{P_1} P(Y_i | X_i).
   \]

Aside: sensible channels not of the above type.

Eg 1. Missing-symbol channel
\[
X_1, X_2, \ldots, X_n \rightarrow Y_1, Y_2, \ldots, Y_m
\]

obtained by \(X_1 \ldots X_n\)
by deleting \(n-m\) symbols.

Don't know which symbols are missing.

Eg 2. Channel
\[
X_1, X_2, \ldots, X_i, X_{i+1}, \ldots, X_n \rightarrow X_1, X_2, \ldots, X_i, X_{i+1}, \ldots, X_n
\]

Symbols can emerge out of order.

Eg 3. Burst errors
\[
X_1, X_2, \ldots, X_n \rightarrow X_1, X_2, \ldots, X_m.
\]

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** → **Code word**
  - 0 → 0 0 ... 0
  - 1 → 1 1 ... 1

- **Decoding** → majority of ⌈k times ⌉

**The code**: set of legitimate code words (a subset of all possible messages).

**Rate**: \( \frac{k}{n} \)

**Prob error**: \( p^k \) for erasure channel

**eg. Hamming code** \( A \times 2 \)

- **Encodes** \( k = 2^r - 1 \) in \( n = 2^r - 1 \) bits

- **Rate**: \( \frac{k}{n} = 1 - \frac{1}{2^r - 1} \)

- Can correct 1 error in BSC.

**eg. Hadamard code**

- Take Hadamard matrix:

- Bit flip error (up to \( \frac{n}{2} - 1 \)) can be corrected. **Rate**: \( \frac{k = \log n}{n} \rightarrow 0 \) as \( n \rightarrow \infty \).
Geometric interpretation.

For simplicity, take $k=1$.

\[ X^m = Y^m \]

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

\[ \text{If few enough codewords} \]
Qn:

1. For a fixed message size, to have smaller & smaller error prob. need bigger & bigger rate
   
   but that means more & more errors.

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

2. For growing message size, how prob. for each segment → 0, can the entire message be sent correctly up to 1?
Sending messages through a noisy channel:

\[ \begin{array}{c}
\text{Sender's lab} \\
\xrightarrow{m} \text{Encoder} \xrightarrow{E_n} \begin{array}{c} y_1 \\ \vdots \\ y_n \end{array} \\
\xrightarrow{\text{Decoder}} \xrightarrow{D_n} \text{Receiver's lab} \\
\end{array} \]

An "\((M, n)\)" code consists of:

1. an index set \(M = \{1, 2, \ldots, M\}\)
2. an encoding function \(E_n : M \rightarrow X^n\)
3. a decoding function \(D_n : Y^n \rightarrow M\)

* The code = \(\{E_n(1), E_n(2), \ldots, E_n(M)\}\)

* Rate of \((M \times n)\) code = \(\frac{1}{n} \log_2 M\).

* For each \(m\), \(m'\) follows a distribution (Randomness introduced by noisy channel).

* Define \(P_e(m) := \text{Prob}(m' \neq m|m)\).

* Define \(P_e := \max_{m \in M} P_e(m)\) \(\text{worse case error}\).

* Define \(\bar{P}_e := \frac{1}{M} \sum_{m \in M} P_e(m)\) \(\text{average error}\).
Def. An achievable rate $J$:

For a channel $N$, a rate $R$ is achievable if a sequence of $(\frac{1}{2}^n R, n)$ codes $C_n$ s.t. $Pe(C_n) \to 0$ as $n \to \infty$.

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

NB: If $C(N) > 0$, the entire message (longer & larger, $2^n R$ bits) comes out correctly (in each symbol) almost surely!

\[
\text{Thm (Shannon's noisy coding theorem)}
\]

\[
C(N) = \max_{p_X} I(X;Y)
\]

where $p_{Y|X}(y|x) = p(y|x)$.

\[
\text{with respect to Channel}
\]

\[
\text{comes in}
\]

NB1. LHS is an asymptotic quantity. No direct relation.

RHS is a "single-letter" formula — involving only 1 copy of $N$.

NB2. Since works in worst case, message need not follow any distribution. (Will see when goes wrong.)
\[ C(n) = 0 \quad \text{iff} \quad p(x, y) = p(x) \cdot p(y|x) = p(x) \cdot p(y). \]

So all other channels have \( C(n) > 0 \).}

\[ I(x:y) = H(y) - H(y|x) \quad \text{for} \quad H(x) \]

Choose \( p(x) \) to max \( H(Y) \)

\[ H(Y|x) = \frac{p(x)}{p(x)} \]

\[ I(x:y) = H(x) - H(x|y) = 1 - p \]

Choose \( p(x) \) to

\[ H(x|1) = 0 \]

\[ H(x|0) = 1 \]

\[ I(x|y) = p \]
Note zero error channel capacity.
How to prove (1)?

1. Prove a converse:
   
   i.e. at any higher rate, $P_e \to 0$

   [Thus $C(N) \leq \max_{p(x)} I(x:y)$]

2. Prove a direct coding theorem:
   
   i.e. given any $p(x)$,

   Show 3 codes achieving the rate $I(x:y)$

   + error analysis $P_e \to 0$

   [Thus $C(N) \geq \max_{p(x)} I(x:y)$]

} holds for completely different reasons.
Direct coding theorem:

- Need to show $\mathbb{P}( \text{E} \cap \text{M} )$ satisfies $\text{C}_n$ with:
  - rate $\frac{1}{n} \log M \geq I(X; Y) - d_n$ as $n \to \infty$
  - error $\mathbb{P}( \text{E} \cap \text{M} ) \to 0$

- Strangely, we need not find these rates.

- Idea: for each $n$, a code $\text{C}_n$ is generated via a random process.

\[
\mathbb{E}_{\text{C}_n} \mathbb{P}( \text{E} \cap \text{M} ) \to 0 \quad \text{while } d_n \to 0
\]

over message

So, some $\tilde{\text{C}}_n$ has $\mathbb{E}_{\tilde{\text{C}}_n} \mathbb{P}( \tilde{\text{E}} \cap \text{M} ) \to 0$.
Derive a matched $\tilde{\text{C}}_n$ with $\mathbb{P}( \tilde{\text{E}} \cap \text{M} ) \to 0$.

Given any $n, M, p(x)$, the random process generating $\text{C}_n$ is:

\[
\text{for } i = 1 \ldots M \quad j = 1 \ldots n
\]

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

Let $\text{C}_n = [x_{11} \quad x_{12} \ldots x_{1n} \quad \ldots \quad x_{M1} \quad x_{M2} \ldots x_{Mn}]$ the $M$ code words defining the code $\text{C}_n$.
Why $P_c(G^n) 	o 0$ if $M < 2^{n(H(Y|X) - \delta_n)}$?

- Output $Y^n$ is i.i.d. $\mathcal{N}(0,1)$
- There are $2^{nH(Y)}$ typical outcomes.

If message is $i$, $C_i$ transmitted via $N^{\omega^n}$
- then there are $2^{nH(Y|X)}$ outcomes, centred at $C_i$, forming a set $S_i$

![Diagram showing sets $S_1, S_2, S_3$]

- If $2^{nH(X|Y)} M \ll 2^{nH(Y)}$
  then $S_1, S_2, \ldots, S_m$ don't overlap much.

Bob
- I 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. \( p(xy) \), drawn \( n \) times,

\( xy^n \) is \( \varepsilon \)-jointly typical if

\[
\begin{align*}
(5) & \quad | \frac{1}{n} \log p(x^n) - H(x) | \leq \varepsilon \\
(6) & \quad | \frac{1}{n} \log p(y^n) - H(y) | \leq \varepsilon \\
(7) & \quad | -\frac{1}{n} \log p(x^n y^n) - H(xy) | \leq \varepsilon
\end{align*}
\]

\( (x^n \text{-typical}) \) \( (y^n \text{-typical}) \) \( (xy^n \text{-typical}) \)

Def: [Jointly typical set]

\( A_{n, \varepsilon} = \{ x^n, y^n : x^n y^n \text{-typical} \} \)

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

Thm: [Joint AEP]

Using above defns. \( \forall \delta > 0 \) \( \forall \varepsilon > 0 \) \( \exists n_0 \) st. \( n \geq n_0 \)

1. \( \Pr \left( A_{n, \varepsilon} \right) \geq 1 - \delta \\
(\text{Hypothesis})

2. \( \left(1 - \delta \right)^n \leq |A_{n, \varepsilon}| \leq 2^n \left( H(xy) + \varepsilon \right) \)

3. If we draw \( (x^n, y^n) \) according to some other dist. \( q \)

\[
q(x^n, y^n) = p(x^n) \cdot p(y^n)
\]

then \( \Pr \left( x^n, y^n \in A_{n, \varepsilon} \right) \leq 2^{-n \left( I(x,y) - 3\varepsilon \right)} \)

\( \Pr = \frac{1}{n} \)
More specifically:

1. A random code $C_n$ is generated.

2. $C_n$ is told to Alice & Bob

3. A message $i$ is drawn uniformly from $f(1, 2, ..., M)$.

4. Alice sends $C_i$ through $N^n$.

5. Bob receives $Y^n \sim Pr(Y^n|C_i)$.

6. Bob decides the outcome $y^n$:

   - If $\exists \ i, j$ st. $y^n \in A_n$ & Bob outputs $j$.

   - Otherwise Bob outputs "failed".
If we make a table:

<table>
<thead>
<tr>
<th>( y^n(1) )</th>
<th>( y^n(2) )</th>
<th>( y^n )</th>
<th>( y^n(2^{H(Y+3)}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x^n(1) )</td>
<td>( x^n(2) )</td>
<td>( x^n )</td>
<td></td>
</tr>
<tr>
<td>( x^n )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n(HX+\epsilon) )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( M \) random rows are code words.

\( T_{\text{m.e.}} \) for X

\[ \text{If } x^n y^n \text{ is ans., but not 1", else 0".} \]

Roughly \( 2^n H(XY) \) 1's
When message is \( \hat{c} \), Alice transmits row \( "c.i." \)

By TAEPC(1) \( wp > 1-d \), the following holds:

Bob receives \( c.i. st. c \in \mathcal{E} \rightarrow \mathcal{A} \).

When Bob sees:

If column for \( c.i. \) restricted to the code, has a unique "1", that "1" is in the row \( c.i. \).

Bob's only but \( j = \hat{i} \).

Only error comes from:

1. TAEPC(1) fails.
2. There is some \( c_k \), \( k \neq j \)

\( \text{st. } c_k, c_i \in \mathcal{A} \in \mathcal{E} \).

But \( c_i \) depends on \( c_k \), indep \( c_k \).

\[ \text{Pr} (c_k (c_i \in \mathcal{A} \cap \mathcal{E}) \leq \text{independent}) \text{ by TAEPC(2).} \]

Over choice \( c_k \).
\[
\frac{1}{n} \sum_{C_n} \mathbb{E} \Pr (C_n) \\
= \frac{1}{n} \sum_{C_n} \mathbb{E} \mathbb{P}_e (C_n) \\
= \frac{1}{n} \sum_{C_n} \mathbb{E} \mathbb{P}_e (C_n) \\
\leq \frac{1}{n} \sum_{C_n} \mathbb{E} \mathbb{P}_e (C_n) \\
\leq d + \sum_{k} \Pr \left( \mathbb{P}_e (C_k \cap C_t^{\prime} \in A_{\eta \varepsilon}) \right) \\
\leq d + M \cdot 2^{-n(\mathbb{H}(X;Y) - 3\varepsilon)} \\
\leq \text{Indef} \eta \cdot n.
\]

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

for large \( n \),

\( M \cdot 2^{-n(\mathbb{H}(X;Y) - 3\varepsilon)} < \eta \)

\( d < \eta \)

\( \mathbb{E} \sum_{C_n} \mathbb{E} \mathbb{P}_e (C_n) < 2\eta \)

So \( \exists C_n \text{ s.t. } \mathbb{E} \mathbb{P}_e (C_n) < 2\eta \).
From $\tilde{E}_n$, get a code $\tilde{C}_n$ with $P_e(\tilde{C}_n) \leq 2^{-n} P_e(\tilde{E}_n)$.

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

$\tilde{E}_n$ codewords, 1 fewer bit than $\tilde{C}_n$.

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

Techniques:

- Random codes
- Symmetry arguments
- Existential proof
- Expanding worse codewords to proof $P_e(\tilde{E}_n)$
Converse:

\[ nR = H(M) = H(M|Y^n) + I(M;Y^n). \]

\[
\begin{align*}
1 & \leq 1 + Pe \cdot nR \\
2 & \leq I(E^n(M);Y^n) \\
3 & \leq n \max_{p(x)} I(x;Y).
\end{align*}
\]

"Fano's inequality:

If \( Pe = 0 \), can deduce \( M \) from \( Y^n \).

"Data processing inequality:

\[ H(\mathbb{E} \hat{X} | Y) \leq H(\mathbb{E} \hat{X} | Y, E^n(M); Y^n) \]

1. Thin Fano's inequality:

Let \( Pe = \text{prob}(X=\hat{X}) \), \( \mathbb{E} = f(Y) \)

\[ \mathcal{S} = \text{sample space of } X. \]

Then \( H(Pe) + Pe \cdot (\log |2^n - 1|) \geq H(X | Y) \).

binary entropy for

Pf: Define new \( rv E \in \mathbb{S} \).

\[ E = 0 \text{ if } X = \hat{X} \]

\[ E = 1 \text{ otherwise.} \]

\[ H(E | X,Y) = H(X,Y) + H(E | X,Y), \]

\[ = H(\mathbb{E} | Y) + H(X | E,Y). \]

\[ = H(\mathbb{E} | Y) + H(X | E = 0, Y | E = 0). \]

Exchange \( \mathbb{E} \) sym on both sides.

\[ H(X | E = 0, Y | E = 0) = H(X | E = 0, \mathbb{E} = 0). \]

\[ \leq H(X | E = 0, Y | E = 0). \]

\[ \leq H(X | E = 0, \mathbb{E} = 0). \]

\[ \leq H(\mathbb{E} = 0) + \sum_y p(y) \left[ Pe \cdot H(X | E = 1, Y | Y) + (1-Pe) \cdot H(X | E = 0, Y | E = 0) \right]. \]

\[ \leq H(\mathbb{E}) + Pe \log (2^{n-1}). \]
Replacing $X$ by $M$

$Y$ by $Y^n$

$\Omega$ by $2^n$

\[ H(M|Y^n) \leq H(P_e) + \log n R \quad \text{apriori} \]

\[ \leq 1 + \log n R. \]

\( \text{(2) Thm 1: Data Processing Inequality (DPI)} \):

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

\[ I(E:B|F) = 0 \]

Then $I(E:F) > I(E:B)$.

\[ \text{Proof:} \quad I(E:F) = H(E) - H(E|F) \]

\[ = I(E:B) + H(E|F) - H(E|F,G) \]

\[ = I(E:B) + I(E:F|G) \]

" LHS sym wrt exchanging $F$ & $G$. "

\[ I(E:F) \geq 0 \quad \text{by SSA} \]

\[ = 0 \quad \text{for } E \rightarrow F \rightarrow G. \]

\[ I(E:F) \leq I(F|G). \]
Lemma: Let $Y^n = N^{\omega^n}(X^n)$

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

Note neither $X^n$ nor $Y^n$ need to be iid.

E.g. $X^n = 111100$

$Y^n = 00 \ldots 00$

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

$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y_{i-1}, \ldots, Y_1, X^n)$ Chainrule

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

$\leq \sum_{i=1}^{n} H(Y_i) - \frac{n}{2} H(Y^n | X^n)$ Source Coding

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

E.g. Back comm doesn't help.