Theorem [Shannon's noisy coding theorem]
\[ C(N) = \max \ p(x) \ I(X:Y) \]

How to prove this?

1. Direct coding – consider codes that are promising

   A clever code doesn't come by easily.
   Instead, consider "random" (M,n) codes with rate = I(X:Y) and show \( \text{Prob}(E_{P_e} \to 0) > 0 \).

   Thus \( \exists \) code with small \( E_{P_e} \) (our 2nd encounter with "existential proofs"). Extract a subcode with similar rate but \( P_e \to 0 \).

2. Converse – show that if at higher rates, \( E_{P_e} \not\to 0 \).

Plan: 2, then heuristic 1, then 1.
Proof of converse:

\[ nR = H(M) = H(M|Y^n) + I(M:Y^n) \]
\[ \leq H(M|Y^n) + I(E_n(M):Y^n) \]

(i) data processing ineq

(ii) \[ H(M|Y^n) \leq 1+P_e nR \]

(iii) by lemma

\[ \leq n \max_{p(x)} I(X:Y) \]
(i) Data processing inequality $I(E:F) \geq I(E:G)$
if $E \rightarrow F \rightarrow G$ is a Markov Chain
(i.e. $I(E:G|F) = 0$)

Proof:

$I(E:FG) = H(E) + H(FG) - H(EFG)$
$= I(E:G) + H(E|G) + H(FG) - H(EFG)$
$= I(E:G) + H(E|G) + - H(E|FG)$
$= I(E:G) + I(E:F|G)$

but the LHS is symmetric wrt exchange $F$ and $G$, so must the RHS.

So, $I(E:G) + I(E:F|G) = I(E:F) + I(E:G|F)$
$\geq 0$ \hspace{2cm} \hspace{2cm} 0

So, $I(E:G) \geq I(E:F)$. 
(ii) Thm [Fanos ineq]:

Let $P_e = \text{prob}(X \neq Z)$, $Z = f(Y)$, $\Omega =$sample space of $X$.
Then, $H(P_e) + P_e \log(|\Omega|-1) \geq H(X|Y)$

Proof: Define new rv $E$, $E=0$ if $X=Z$, $E=1$ otherwise.

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

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

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

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

Making the replacements:
$M \leftrightarrow X$
$Y^n \leftrightarrow Y$
$2^{nR} \leftrightarrow |\Omega|$ gives $H(M|Y^n) \leq 1+P_e nR$
(iii) Lemma: Let \( Y^n = N^{\otimes n}(X^n) \).

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

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

\[= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1 \ldots Y_{i-1}X^n) \quad \text{Chain rule} \]

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

\[\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{Subadditivity} \]

\[\leq \sum_{i=1}^{n} I(X_i:Y_i) \]
1. Direct coding:

Let \( M = 2^{n(I(X:Y)-\delta n)} \). What's the \((M,n)\) code?

Fix any \( p(x) \).

Encoder \( \mathcal{E}_n \):
Pick \( M \) codewords \( c_i = x_{i1} \ldots x_{in} \)

\( \text{each } x_{ij} \text{ chosen iid } \sim p(x) \).

Fixed & known to Alice & Bob once choosen.

\( c_1 = x_{11}, x_{12}, \ldots, x_{1n} \)
\( c_2 = x_{21}, x_{22}, \ldots, x_{2n} \)
\( \cdots \)
\( c_M = x_{M1}, x_{M2}, \ldots, x_{Mn} \)

Everything refers to this particular code \( \mathcal{C}_n \) from now on.
1. Direct coding:

\[ c_1 = x_{11}, x_{12}, \ldots, x_{1n} \]
\[ c_2 = x_{21}, x_{22}, \ldots, x_{2n} \]
\[ \vdots \]
\[ c_M = x_{M1}, x_{M2}, \ldots, x_{Mn} \]

\[ x_{ij} \text{ chosen iid } \sim p(x) \]

Heuristically why \( P_e \to 0 \):

The \( n \) channel outputs \( Y^n \) is iid with \( p(y) = \sum_x p(y|x) \ p(x) \)

With high prob, output typical \( y_1 \cdots y_n, \approx 2^{nH(Y)} \) of them.

For each \( c_i \) sent via \( N^\otimes n \), there're \( \approx 2^{nH(Y|X)} \) possible outcomes (call the set \( O_i \)) centered around \( c_i \).

Since the \( c_i \)'s are random, if \( 2^{nH(Y|X)} M \ll 2^{nH(Y)} \), these \( O_i \)'s don't overlap much. So, decoder just output "which sphere" contains the output \( y_1 \cdots y_n \).
1. Direct coding: 
\[ c_1 = x_{11}, x_{12}, \ldots, x_{1n} \]
\[ c_2 = x_{21}, x_{22}, \ldots, x_{2n} \]
\[ \vdots \]
\[ c_M = x_{M1}, x_{M2}, \ldots, x_{Mn} \]

\( x_{ij} \) chosen iid \( \sim p(x) \)

Better proof why \( P_e \to 0 \).

Long version (18 pages) available in homepage, here, shrink to 6 pages, skipping most detail, esp \( \varepsilon, \delta \) ignored.
Recall:

**Def[typical sequence]:**
$x^n \epsilon$-typical if $|-1/n \log(p(x^n)) - H(X)| \leq \epsilon$
It means $2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)}$.

**Def[Jointly typical sequence]:**
$x^n y^n \epsilon$-jointly-typical if
$|-1/n \log(p(x^n y^n)) - H(XY)| \leq \epsilon$
where $p(x^n y^n) = \Pi_{i=1}^{n} p(x_i y_i)$.

Need also: (a) $|-1/n \log(p(x^n)) - H(X)| \leq \epsilon$  
(b) $|-1/n \log(p(y^n)) - H(Y)| \leq \epsilon$  
[The strong typicality has (c) $\Rightarrow$ (a,b), but not for entropic typicality.]

**Def[Jointly-typical set]:** $A_{n,\epsilon} = \{x^n y^n \epsilon$-jointly typical$\}$
Joint asymptotic equipartition (Joint AEP) theorem:

Let \((X^n,Y^n)\) be sequences of length \(n\)
drawn iid according to \(p(x^n y^n) = \prod_{i=1}^{n} p(x_i y_i)\).

Then:
1. \(\Pr(X^nY^n \in A_{n,\varepsilon}) \to 1\)
2. \(|A_{n,\varepsilon}| \approx 2^{nH(XY)}\)
3. if we draw \(X^n\) & \(Y^n\) according to \(q(x^n y^n) = p(x^n) p(y^n)\).
   \(\Pr_q (\text{outcome } \in A_{n,\varepsilon}) \approx 2^{-nI(X:Y)}\)

Proof (with \(\varepsilon, \delta\)) available in the 18 page notes.
More observations:

Given \( y^n \in T^Y_{n,\varepsilon} \), how many \( x^n \in T^X_{n,\varepsilon} \) is s.t. \( x^n y^n \in A_{n,\varepsilon} \)? Call this set \( S(y^n) \).

(1) \[ p(x^n|y^n) = p(x^n y^n) / p(y^n) \approx 2^{-n[H(XY)-H(Y)]} = 2^{-n[H(X|Y)]} \]
\[ \uparrow \text{since } x^n y^n \in A_{n,\varepsilon} !! \]

(2) \[ 1 = \sum_{x^n \in S} p(x^n|y^n) \approx |S(y^n)| \cdot 2^{-n[H(X|Y)]} \]

Hence, \(|S(y^n)| \approx 2^{nH(X|Y)}\). Fraction of such \( x^n \approx 2^{-nI(X:Y)} \).

Similarly, given \( x^n \in T^X_{n,\varepsilon} \), \( \approx 2^{nH(Y|X)} \) \( y^n \)'s are jointly typical with it, and the fraction of such \( y^n \approx 2^{-nI(X:Y)} \).
Make a table of typical $x^n$'s and $y^n$'s, and for jointly typical $x^n y^n$, put a 1, else, put a 0.

<table>
<thead>
<tr>
<th></th>
<th>$y^n(1)$</th>
<th>...</th>
<th>$y^n(\ )$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^n(1)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x^n(2)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x^n(\ )$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each row has $\approx 2^{nH(Y|X)}$ 1's.
Each column $\approx 2^{nH(X|Y)}$ 1's.
Total: $\approx 2^{nH(XY)}$ 1's.

Our random code corresponds to $M$ randomly chosen rows.
D_n: typical set decoding

Given y^n, if there is a unique x^n ∈ S(y^n),
    output m' s.t. c_m = x^n.
Else, output W=M+1 (error symbol).

How will this fail for message m?
Either - no such x^n
    - or ∃m"≠m with c_{m"}y^n ∈ A_{ε,n}

For the random code C_n, let the average error over all messages
be EP_e(C_n), same as error if m=1 (since all messages similar).

\[
EP_e(C_n) = \Pr_{C_n}(W\neq 1|m=1) = \Pr_{C_n}(\text{Err}_0 \cup \text{Err}_2 \ldots \cup \text{Err}_M |m=1)
\]
union bdd \(\leq M \Pr_{C_n}(\text{Err}_2 |m=1)\) equiprobable
Bounding $\Pr_{c_n}(\text{Err}_2|m=1) = \Pr_{c_n}(c_2y^n \in A_{n,\varepsilon_n})$:

But $c_2$ and $y^n = N^{\otimes n}(x_1)$ independent.

By joint AEP \cite{3}, $\Pr_{c_n}(c_2y^n \in A_{n,\varepsilon_n}) \approx 2^{-nI(X:Y)}$

If $M = 2^n(I(X:Y) - \delta_n)$ and $n\delta_n$ growing with $n$ but $\delta_n \to 0$

$$\operatorname{EP}_e(C_n) \leq M \Pr_{c_n}(\text{Err}_2|m=1) \to 0$$

Note: $\operatorname{Pe}(m=1) + \operatorname{Pe}(m=2) + \ldots + \operatorname{Pe}(m=M) = M \operatorname{EP}_e(C_n)$

Reorder $m$'s so that $\operatorname{Pe}(m)$ is increasing.

So, $\operatorname{Pe}(m=1) + \operatorname{Pe}(m=2) + \ldots + \operatorname{Pe}(m=M/2) \leq M/2 \operatorname{EP}_e(C_n)$

So, keeping only codewords for $m=1, \ldots, M/2$, worse case prob error $\leq \operatorname{EP}_e(C_n)/2$. 