Definition: The "rate" of a \((M, n)\) code is \(R := \frac{\log M}{n}\)

(the unit are \(\text{bits/channel use}\) when we use the binary logarithm).

- \(R \in \mathbb{R}_+\) is called an "achievable rate" for a given channel if there exist a sequence of \(\left(\frac{1}{2^R}, n\right)\) codes, with \(\lambda^{(n)} \to 0\) as \(n \to \infty\).

- The capacity \(C(T)\) of a channel \(T\) is the supremum over all achievable rates.

Example: The \(n\)-fold repetition code \(0 \mapsto 0^n := (0, \ldots, 0) \in \{0, 1\}^n\)

has rate \(R = \frac{\log 2}{n} = \frac{1}{n}\) and (for odd \(n\))

\[\lambda^{(n)} \geq \frac{1}{n}\] (for binary symmetric channel)

Thus, this sequence of channels has \(R \to 0\) if we require \(\lambda^{(n)} \to 0\).

I.8 Shannon's noisy channel coding theorem

Definition (jointly typical set):

Let \(n \in \mathbb{N}, \varepsilon > 0\), and let \(p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+\) be the joint probability distribution of random variables \(X\) and \(Y\). The set of "jointly typical sequence\) with respect to the joint distribution \(p\) is defined as

\[B_{\varepsilon}^{(n)} := \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \mid \left| -\frac{1}{n} \log p(x) - H(X) \right| < \varepsilon \right.\]

and \(\left| -\frac{1}{n} \log p(y) - H(Y) \right| < \varepsilon\)

and \(\left| -\frac{1}{n} \log p(xy) - H(XY) \right| < \varepsilon\}\)

where \(p(x,y) = \prod_{i=1}^{n} p(x_i, y_i)\) and \(p(x) = \prod_{i=1}^{n} p(x_i)\), \(p(y) = \prod_{i=1}^{n} p(y_i)\)

are the marginals.
Theorem (Joint AEP):

Let $B^{(n)}_\varepsilon$ be the set of jointly typical sequences w.r.t. the joint distribution of $X$ and $Y$. Then:

1. $P(B^{(n)}_\varepsilon) > 1 - \varepsilon$ for all sufficiently large $n$.
2. $|B^{(n)}_\varepsilon| \leq 2^{-n(H(X)+\varepsilon)} \forall n$.
3. $|B^{(n)}_\varepsilon| \geq (1-\varepsilon) 2^{-n(H(X)-\varepsilon)}$ for all sufficiently large $n$.
4. If $\tilde{X}^n, \tilde{Y}^n$ are random variables with

$$
P(\tilde{X}^n = (x_1, \ldots, x_n), \tilde{Y}^n = (y_1, \ldots, y_n)) = \prod_{i=1}^{n} p(x_i) p(y_i)$$

[i.e. the i.i.d. variables $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginal, as $p(x)$ and $p(y)$], then

$$
P_r[(\tilde{X}^n, \tilde{Y}^n) \in B^{(n)}_\varepsilon] \leq 2^{-n(I(X;Y)-3\varepsilon)} \forall n,$$

and

$$
P_r[(\tilde{X}^n, \tilde{Y}^n) \in B^{(n)}_\varepsilon] \geq (1-\varepsilon) 2^{-n(I(X;Y)+3\varepsilon)}$$

for all sufficiently large $n$.

Proof: (1), (2), (3) follow very similarly as the AEP for $A^{(n)}_\varepsilon$. For (4), note that each of the three terms, like

$$
P((X^n, Y^n) \in B^{(n)}_\varepsilon) = P\left[\left| -\frac{1}{n} \log p(X^n, Y^n) - H(XY) \right| > \varepsilon \right] \to 0 \text{ as } n \to \infty,$$

converge to 0 individually as $n \to \infty$; i.e. each term is smaller than $\varepsilon/3$ for sufficiently large $n$.

(4) $P_r[(\tilde{X}^n, \tilde{Y}^n) \in B^{(n)}_\varepsilon] = \sum_{(x,y) \in B^{(n)}_\varepsilon} p(x) p(y)$

$$
\leq |B^{(n)}_\varepsilon| 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)}
$$

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

$$
\leq 2^{-n(H(X))+2^{-n(H(Y))+2^{-n(I(X;Y)-3\varepsilon}}}
$$
and \( P = \{ (X^n, Y^n) \in \mathcal{B}^{(n)}_\varepsilon \} = \sum_{(X^n, Y^n) \in \mathcal{B}^{(n)}_\varepsilon} p(x^n) p(y^n) \times |\mathcal{B}^{(n)}_\varepsilon| 2^{-n(H(X)+\varepsilon)} 2^{-n(H(Y)+\varepsilon)} \geq (1-\varepsilon) 2^{-n(I(X;Y)+3\varepsilon)}. \)

We are now in the position to show the achievable rate of some range via "jointly typical decoding".

**Theorem (Achievability of Noisy Channel Coding)**: 

For a discrete memoryless channel \( T : \mathbb{R}^X \rightarrow \mathbb{R}^Y, \) 

every \( R < I(X;Y) \) is an achievable rate, \( \{p(x^n)\} \) 

where the mutual information \( I(X;Y) \) is computed w.r.t. the joint distribution \( p(x,y) = p(x) p(y|T(x)) = p(x) \mathcal{T}_{y|x} (x \in X, y \in Y) \) and the maximum runs over all probability distributions \( p(x) \) on \( X \).

**Proof**: Fix \( p(x) \) such that \( R < I(X;Y) \), let \( M = \{1, 2, \ldots, 2^{|X|} \} \), whose \( n \) will be any sufficiently large integer later, and \( \varepsilon > 0 \) will be chosen appropriately later.

Generate a \( (2^{|X|}, n) \)-code \( C \) randomly by generating the \( 2^{|X|} \) codewords \( x(m) \in X^n \) (for \( m \in M \)) independently, factoring according to the distribution \( p(x^n) = \prod_{i=1}^n p(x_i) \) for \( x \in X^n \).

Use the "jointly typical decoding" \( g : Y^n \rightarrow M \), defined by

\[
g(y^n) = \begin{cases} 
m & \text{if } (x(m), y^n) \in \mathcal{B}^{(n)}_\varepsilon \text{ and } \forall m' \neq m : (x(m'), y^n) \notin \mathcal{B}^{(n)}_\varepsilon \\
1 & \text{otherwise} \end{cases}
\]
We now aim to compute the error probability $p_e^{(n)}$ or $\lambda^{(n)}$ of this code $C$. It will however be easier to compute average error probability $p_e^{(n)}(C)$ averaged over all codes $C$ from above first:

\[
\hat{p} = \sum_C p(e) p_e^{(n)}(C) = \sum_C p(e) 2^{-nK} \sum_m \lambda_m^{(n)}(C)
\]

\[
= 2^{-nR} \sum_m \sum_C p(e) \lambda_m^{(n)}(C) = \sum_C p(e) \lambda_n^{(n)}(C)
\]

Two types of decoding error can occur:

(i) $(X^{(n)}, Y^{(n)}) \notin B^e_m$: this has probability $< \varepsilon$ (joint AEP)

(ii) $(X^{(m')}, Y^{(n)}) \in B^e_m$ for some $m' + 1$:

$x^{(m')}$ and $y^{(n)}$ are independent if $m' + 1$ and $x^{(m')} \sim X^{(n)}$, $y^{(n)} \sim Y^{(n)}$.

Thus, item (4) of joint AEP gives $\Pr[(x^{(m')}, y^{(n)}) \in B^e_m] \leq 2^{-n(I(X;Y) - 3\varepsilon)}$ for $m' + 1$.

\[
\hat{p} < \varepsilon + \left( \frac{2^{-nR} - 1}{m' + 1} \right) \leq 2^{-n(I(X;Y) - R - 3\varepsilon)}
\]

\[
\leq 2\varepsilon \quad \text{if } R < I(X;Y) - 3\varepsilon \text{ and } n \text{ sufficiently large.}
\]