

Channel coding

Shannon's noisy channel coding theorem

Jacob Louis Hoover

2020.12.16

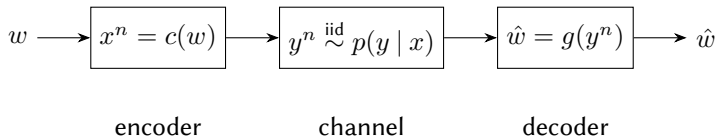
Math 740 — Prof. Vojkan Jakšić

Outline

- 1 background
 - informal motivation
 - preliminaries
 - discrete channel
 - block code
- 2 AEP and typicality
 - (jointly) typical sequences
 - joint AEP
- 3 noisy-channel coding theorem
 - typical-set coding scheme
 - asymptotically optimal code

Suppose information source is perturbed during transmission between encoder and decoder.

In finite case, where perturbation of successive transmissions is independent, this 'noisy channel' can be described as a set of conditional probability distributions over input symbols given output symbols.



$p(y | x)$ probability of transmitted symbol x being received as y .

- Two statistical processes at work: source and noise
- Problem: noise leads to some probability of error (that $\hat{w} \neq w$).
Choose coding scheme to communicate effectively as possible despite this problem.
- Idea: Agree on 'widely spaced' inputs, so that the probability of error is small.

Noisy-channel coding preview

To prove that a good coding exists, calculate average probability of error. Show that average is small, therefore there must exist individual codes with small probability of error.

- for each (typical) X^n , there are $\approx 2^{nH(Y|X)}$ possible Y^n
- Total number of (typical) Y^n is $\approx 2^{nH(Y)}$
- Total number of messages (distinguishable inputs) should be $2^{n(H(Y)/2^{H(Y|X)})} = 2^{nI(X;Y)}$
- To formalize these ideas, we need to discuss **joint typical sequences**

Let X be a discrete random variable on \mathcal{X} , and let p_X be the probability distribution of X .

- entropy

$$H(X) = \mathbb{E}_{p_X}[-\log p_X(x)], \quad H(X, Y) = \mathbb{E}_{p_{X,Y}}[-\log p_{X,Y}(x, y)]$$

- conditional entropy

$$H_Y(X) = \mathbb{E}_{p_{X,Y}}[-\log(\frac{p_{X,Y}(x, y)}{p_Y(y)})] = H(X, Y) - H(Y)$$

- mutual information

$$I(X : Y) = \mathbb{E}_{p_{X,Y}}[-\log(\frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)})] = H(X) - H_Y(X)$$

Definition (discrete channel)

A discrete channel is a tuple $(\mathcal{X}, \{p_x\}_{x \in \mathcal{X}}, \mathcal{Y})$, where \mathcal{X}, \mathcal{Y} are finite alphabets, and for each $x \in \mathcal{X}$, p_x is a probability over \mathcal{Y} .

Interpretation:

- \mathcal{X} is the **input alphabet**
- \mathcal{Y} is the **output alphabet**
- the probabilities $\{p_x\}$ define a transition matrix expressing the probability of observing symbol $y \in \mathcal{Y}$, given symbol $x \in \mathcal{X}$ was sent.

$$M_{xy} = p_x(y) = p(Y = y \mid X = x)$$

where X, Y will always represent random variables over \mathcal{X} and \mathcal{Y} , resp. We can write channel as $(\mathcal{X}, p_{Y|X}, \mathcal{Y})$.

This discussion focuses solely on discrete channels which are memoryless, and without feedback.

Definition

The n th extension of a discrete channel is the channel $(\mathcal{X}^n, \{p_{x^n}\}_{x^n \in \mathcal{X}^n}, \mathcal{Y}^n)$. Say this channel is

- **memoryless** iff $p(y_k | x_1^k, y_1^{k-1}) = p(y_k | x_k)$ for all $1 \leq k \leq n$; and
- **without feedback** iff $p(x_k | x_1^{k-1}, y_1^{k-1}) = p(x_k | x_1^{k-1})$ for all $1 \leq k \leq n$.

The transition function for a discrete memoryless channel without feedback factorizes as

$$p_{Y^n|X^n} = \prod_{i=1}^n p_{Y|X},$$

so channel is specified by a pointwise transitions $(\mathcal{X}^n, p_{Y|X}, \mathcal{Y}^n)$

Definition (channel capacity)

The information channel capacity of a channel $(\mathcal{X}, \{p_a\}_{a \in \mathcal{X}}, \mathcal{Y})$ is

$$C = \max_{p_X \in \mathcal{P}(\mathcal{X})} I(X : Y) = \max_{p_X \in \mathcal{P}(\mathcal{X})} H(X) - H_Y(X)$$

Shannon calls this conditional entropy the ‘equivocation’ – average ambiguity of the received signal.

An $(M \in \mathbb{N}^+, n \in \mathbb{N}^+)$ coding scheme encodes M different messages from the source into codewords in \mathcal{X}^n . WLOG, Let the messages simply be the integers $1, \dots, M$.

Definition (block code)

A (M, n) code consists of an encoding function $c : \{1, \dots, M\} \rightarrow \mathcal{X}^n$, and a decoding function $g : \mathcal{Y}^n \rightarrow \{1, \dots, M\}$

- The *rate* of a (M, n) code is $R = \frac{1}{n} \log M$ (bits per transmission).
- Denote by $\lambda_w^{(n)}$ the conditional probability of error:

$$\lambda_w^{(n)} = \Pr(g(Y^n) \neq w \mid X^n = c(w)) = \sum_{\{y^n : g(y^n) \neq w\}} p(y^n \mid c(w))$$
- Denote $\lambda_{\max}^{(n)} := \max_w \lambda_w^{(n)}$
- A rate R is *achievable* if there exists sequence of $(2^{nR}, n)$ codes such that the maximal conditional error vanishes for large enough n .

$$\lambda_{\max}^{(n)} \rightarrow 0 \quad (n \rightarrow \infty)$$

- Denote by $\lambda_{\text{mean}}^{(n)}$ the mean conditional probability of error, over all codewords, $\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w^{(n)}$.

Theorem (AEP)

For an x_1^n iid sequence drawn from p_X , $\frac{-1}{n} \log p_{X^n}(x_1^n) \xrightarrow{p} H(X)$.

- Proof is immediate from (weak) law of large numbers.

Define **typical sequences** as those for which the empirical entropy is close to the true entropy.

Definition (typicality)

A sequence x_1^n is *typical* of distribution p_X , to tolerance ϵ if

$$2^{-n(H(X)+\epsilon)} \leq p_{X^n}(x_1^n) \leq 2^{-n(H(X)-\epsilon)}.$$

Denote by $A_\epsilon^{(n)}(p_X)$ the ϵ -*typical set* with respect to p .

AEP has consequences for the typical set, for its

- measure, $\Pr\{X^n \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$ (justifies ‘typical’)
- size, $|A_\epsilon| \approx 2^{nH(X)}$ for large n , (lower entropy means smaller typical set)

Definition (joint typicality)

A pair of sequences (x_1^n, y_1^n) are *jointly typical* of distribution p_{XY} (to tolerance ϵ) if all three of the following requirements hold:

- x^n typical of p_X : $\left| \frac{-1}{n} \log p_X(x_1^n) - H(X) \right| < \epsilon$
- y^n typical of p_Y : $\left| \frac{-1}{n} \log p_Y(y_1^n) - H(Y) \right| < \epsilon$
- (x^n, y^n) typical of p_{XY} : $\left| \frac{-1}{n} \log p_{X^n Y^n}(x_1^n, y_1^n) - H(X, Y) \right| < \epsilon$
 where $p_{X^n Y^n} = \prod p_{XY}$

Denote by $A_\epsilon^{(n)}(p_{XY})$ the ϵ -*typical set* with respect to p_{XY} .

We will use the following consequences of AEP and jointly typicality in designing the decoder later.

Consequences of AEP for joint typical sequences:

Let (X^n, Y^n) be drawn iid from p_{XY}

1. probability that X^n, Y^n are ϵ -jointly typical $\rightarrow 1$.
2. The size of the jointly typical set is close to $2^{nH(X,Y)}$
3. If \tilde{X}^n and \tilde{Y}^n are independent samples with distributions identical to the marginals of p_{XY} , then the probability of their being jointly typical is close to $2^{-nI(X:Y)}$.

Theorem (joint AEP)

Let $X^n, Y^n \sim p_{X^n Y^n}(x^n, y^n) = \prod_i p_{XY}(x_i, y_i)$. Then

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
3. If $(\tilde{X}^n, \tilde{Y}^n) \stackrel{iid}{\sim} p_X \cdot p_Y$, then $\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon\} \leq 2^{-n(I(X:Y)-3\epsilon)}$

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$

Proof.

By AEP, for $\epsilon > 0$ there are n_1, n_2, n_3 such that

- for $n \geq n_1$, $\Pr\{|-\frac{1}{n} \log p_{X^n}(x^n) - H(X)| \geq \epsilon\} < \epsilon/3$
- for $n \geq n_2$, $\Pr\{|-\frac{1}{n} \log p_{Y^n}(y^n) - H(Y)| \geq \epsilon\} < \epsilon/3$
- for $n \geq n_3$, $\Pr\{|-\frac{1}{n} \log p_{X^n Y^n}(x^n, y^n) - H(X, Y)| \geq \epsilon\} < \epsilon/3$

Choose n greater than $\max\{n_1, n_2, n_3\}$.

Let U be union of sets above. $\Pr(U) < \epsilon$.

So, for large enough n , $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$. □

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$

Proof.

Within the typical set, by definition,

$$H(X, Y) - \epsilon \leq \frac{-1}{n} \log p_{X^n Y^n}(x_1^n, y_1^n) \leq H(X, Y) + \epsilon$$

$$2^{-n(H(X,Y)+\epsilon)} \leq p_{X^n Y^n}(x_1^n, y_1^n) \leq 2^{-n(H(X,Y)-\epsilon)}$$

So $|A_\epsilon^{(n)}| 2^{-n(H(X,Y)+\epsilon)} \leq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \leq \sum p(x^n, y^n) = 1$ □

Note also that since

$$|A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n)$$

(for sufficiently large n ,) $\geq 1 - \epsilon$ (by 1.)

so for large n have also $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)}$

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(\mathcal{H}(X,Y)+\epsilon)}$, & for large n , $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(\mathcal{H}(X,Y)-\epsilon)}$
3. If $(\tilde{X}^n, \tilde{Y}^n) \stackrel{\text{iid}}{\sim} p_X \cdot p_Y$, then $\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon\} \leq 2^{-n(\mathcal{I}(X:Y)-3\epsilon)}$

Proof.

$$\begin{aligned}
 \Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} &= \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p_{X^n}(x^n) p_{Y^n}(y^n) \\
 &\leq |A_\epsilon^{(n)}| 2^{-n(\mathcal{H}(X)-\epsilon)} 2^{-n(\mathcal{H}(Y)-\epsilon)} \\
 &\leq 2^{n(\mathcal{H}(X,Y)+\epsilon)} 2^{-n(\mathcal{H}(X)-\epsilon)} 2^{-n(\mathcal{H}(Y)-\epsilon)} \\
 &= 2^{-n(\mathcal{I}(X,Y)-3\epsilon)} \quad \square
 \end{aligned}$$

Similarly to before, for large enough n ,

$$\begin{aligned}
 \Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} &\geq |A_\epsilon^{(n)}| 2^{-n(\mathcal{H}(X)+\epsilon)} 2^{-n(\mathcal{H}(Y)+\epsilon)} \\
 &\geq (1 - \epsilon) 2^{n(\mathcal{H}(X,Y)-\epsilon)} 2^{-n(\mathcal{H}(X)+\epsilon)} 2^{-n(\mathcal{H}(Y)+\epsilon)} \\
 &= (1 - \epsilon) 2^{-n(\mathcal{I}(X,Y)+3\epsilon)}
 \end{aligned}$$

1. $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} \rightarrow 1$ as $n \rightarrow \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$, & for large n , $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)}$
3. If $(\tilde{X}^n, \tilde{Y}^n) \stackrel{\text{iid}}{\sim} p_X \cdot p_Y$, then $\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon\} \leq 2^{-n(I(X:Y)-3\epsilon)}$,
& for large n , $\Pr\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \geq (1 - \epsilon)2^{-n(I(X,Y)+3\epsilon)}$

Remark

There are about $2^{nH(X)}$ typical X sequences, and likewise for Y . Only $2^{nH(X,Y)}$ pairs of sequences are jointly typical, though.

Mutual information of X and Y tells how likely a randomly chosen pair of sequences is to be jointly typical.

This is connected to the intuition behind the typical set decoding scheme.

Theorem (noisy-channel coding¹)

Take any discrete memoryless channel Q , with capacity

$$C = \max_{\mathcal{P}_X} I(X : Y).$$

For any any $\epsilon > 0$ and $R < C$, for large enough n there exists a code for Q of length n and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

For proof,

- To prove existence of a good code, show that average probability of error for a **random choice of code** is small, so there must be at least one individual code with small error probability.
- Allow error, but arbitrarily small, for large enough block length.
- Make use of joint typicality for decoding that is simple to analyze (not an optimal scheme, but sufficient).

¹Shannon (1948, Th. 11); MacKay (2003, Th. 10.1); Cover and Thomas (2006, Th. 7.1.1)

Fix p_X arbitrarily for now. For some rate R' , generate a $(2^{nR'}, n)$ code, with encoding function set by sampling independently from p_X .

$$\begin{aligned}c(1) &= (c(1)_1, c(1)_2, \dots, c(1)_n) \\c(2) &= (c(2)_1, c(2)_2, \dots, c(2)_n) \\&\vdots \\c(2^{nR'}) &= (c(2^{nR'})_1, c(2^{nR'})_2, \dots, c(2^{nR'})_n)\end{aligned}$$

That is, each entry in each codeword is independently generated: $c(w)_i \stackrel{\text{iid}}{\sim} p_X$. Let c stand for this codebook.

$$\Pr(c) = \prod_{w=1}^{2^{nR'}} \prod_{i=1}^n p_X(c(w)_i)$$

1. random code generated according to p_X
2. code revealed to both parties (both know channel transition $p_{Y|X}$)
3. message w chosen uniformly from $\{1, \dots, 2^{nR'}\}$
4. transmission $c(w)$ sent (n uses of the channel)
5. receiver receives sequence y_1^n according to $\prod_{i=1}^n p_{Y|X}(y_i | x_i)$
6. receiver decodes by **jointly typical** decoding:
 - if $\exists \hat{w}$ such that $(c(\hat{w}), y_1^n) \in A_\epsilon^{(n)}$,
and $\nexists w'$ such that $(c(w'), y_1^n) \in A_\epsilon^{(n)}$, decode $g(y_1^n) = \hat{w}$
 - else, declare failure (can say $g(y_1^n) = 0$, or undefined)
7. there is a decoding error if $\hat{w} \neq w$. Let \mathcal{E} be event of error.
 - average probability of error, over all codewords and all codebooks,

$$\hat{p}_{\mathcal{E}} = \sum_c \Pr(c) \frac{1}{2^{nR'}} \sum_{w=1}^{2^{nR'}} \lambda_w^{(n)}(c) = \sum_c \Pr(c) \lambda_1^{(n)}(c)$$

is the avg probability of error of any particular codeword over all codebooks. WLOG choose $w = 1$.

Let E_v denote the event that $(c(v), y_1^n) \in A_\gamma^{(n)}$

(Given that $w = 1$) there are two possible sources of error:

- received seq. *not* typical with transmitted: $(c(1), y_1^n) \notin A_\gamma^{(n)}$
For large enough n , $\Pr(E_1^c) < \delta$ (by Joint AEP1, $\Pr(E_1) \rightarrow 1$).
- wrong* codeword is typical with transmitted: $(c(w'), y_1^n) \in A_\gamma^{(n)}$ for $w' \neq 1$
For any message v , $c(v) \stackrel{\text{iid}}{\sim} p_X$, and also $y_1^n \stackrel{\text{iid}}{\sim} p_Y$.
So, $\Pr(E_v) \leq 2^{-n(I(X:Y)-3\gamma)}$ (by Joint AEP3).

$$\begin{aligned} \hat{p}_E &= \Pr(E_1^c \cup \bigcup_{w' \neq 1} E_{w'}) \leq \Pr(E_1^c) + \sum_{w' \neq 1} \Pr(E_{w'}) \\ &\leq \delta + \sum_{w'=2}^{2^{nR'}} 2^{-n(I(X:Y)-3\gamma)} = \delta + (2^{nR'} - 1)2^{-n(I(X:Y)-3\gamma)} \\ &\leq \delta + 2^{-n(I(X:Y)-3\gamma-R)} \end{aligned}$$

$\hat{p}_E \leq 2\delta$ for sufficiently large n , if $R' < I(X : Y) - 3\gamma$. Thus for any rate $R' < I(X : Y)$, we can take n large enough that $\hat{p}_E \leq 2\delta$.

We have: if rate $R' < I(X : Y)$ then $\hat{p}_{\mathcal{E}} \leq 2\delta$ for large enough n .

We need: a *particular* code which has vanishing *maximum* probability of error, for rate below capacity, C .

1. Fix $p_X = \arg \max_{\mathcal{P}(X)} I(X : Y)$, the distribution which achieves capacity.

Thus, rate requirement becomes $R' < C$

2. $\hat{p}_{\mathcal{E}} \leq 2\delta$, so some code has mean probability of block error $\leq 2\delta$.

$$\exists c^* \lambda_{\text{mean}}^{(n)}(c^*) = \frac{1}{2^{nR'}} \sum_{w=1}^{2^{nR'}} \lambda_w^{(n)}(c^*) \leq 2\delta$$

3. Delete² worst half of the codewords. New codebook c' of remaining $2^{nR'} - 1$ codewords achieves maximal block error $\lambda_{\text{max}}^{(n)}(c') \leq 4\delta$.
[Note, rate of this code is reduced, from R' to $R' - \frac{1}{n}$]

²Process has fancy name, 'expurgation' (MacKay, 2003, p. 167)

We constructed code with rate $R' - \frac{1}{n}$, for $R' < C$, with maximal probability of error $< 4\delta$.





So, any rate below the channel capacity is achievable.

Remark

The code c' above achieves the rate $R < C$ for a given ϵ (as in the statement of the theorem), by the above argument, setting

- $R' = (R + C)/2$,
- $\delta = \epsilon/4$,
- $\gamma < (C - R)/3$,
- *and n sufficiently large*



-  Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. USA: Wiley-Interscience. doi: 10.1002/047174882X.
-  MacKay, David J. C. (2003). *Information theory, Inference and Learning Algorithms*. Cambridge university press. eprint: <http://www.inference.org.uk/itprnn/book.pdf>.
-  Shannon, Claude E. (1948). “A mathematical theory of communication”. In: *Bell Syst. Tech. J.* 27.3, pp. 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.
-  Shields, Paul C. (1996). *The ergodic theory of discrete sample paths*. Vol. 13. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI. doi: 10.1090/gsm/013.