Information Theory Notes
Cover & Thomas Chapters 2-5, 7-9

David Abel*
david_abel@brown.edu

I made these notes while taking APMA 1710 at Brown during Fall 2016 (taught by Prof. Govind Menon[1]), which followed the 2nd edition of the Cover & Thomas Information Theory Textbook [1]. If you find typos, please let me know at the email above. The images are of course based on the textbook, but are of my own creation.

Contents

*http://david-abel.github.io
[1] http://www.dam.brown.edu/people/menon/

1 Chapter Two: Entropy

1.1 Definitions

Entropy: $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$

Joint Entropy: $H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$

Conditional: $H(X \mid Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x \mid y)$

Relative Entropy: $D(p \mid\mid q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$

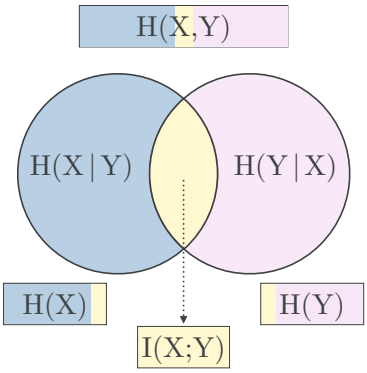
Mutual Information: $I(X;Y) = D(p(x,y) \mid\mid p(x)p(y)) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

Others: $I(X;Y \mid Z), I(X_1, \ldots, X_n; Y \mid Z), H(X_1, \ldots, X_n \mid Z), H(X,Y \mid Z), D(p(y \mid x) \mid\mid q(y \mid x))$

$$H(X,Y \mid Z) = H(X \mid Z) + H(Y \mid X,Z)$$
$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$$

Note: can get at these from a three-way venn diagram

1.2 Identities



Bounds:

- $0 \leq H(X) \leq_U \log |\mathcal{X}|$, where $\leq_U$ is with equality iff $p(x)$ is the uniform distribution.
- $0 \leq H(Y \mid X) \leq_I H(Y)$, where $\leq_I$ is with equality iff X and Y are independent.
- $0 \leq H(X,Y) \leq_I H(X) + H(Y)$, where $\leq_I$ is with equality iff X and Y are independent.
- $I(X;X) = H(X)$
- $0 \leq I(X;Y) \leq H(X)$
- $D(p \mid\mid q) \geq 0$, with equality only if $p = q$.

1.3 Convexity

Convexity: for $\lambda \in [0, 1]$:

$$\underset{①}{f(\lambda x_1 + (1-\lambda)x_2)} \leq \lambda f(x_1) + \underset{②}{(1-\lambda)f(x_2)} \tag{1}$$

Where ① specifies any point on the function between x_1 and x_2, and ② specifies any point on the line connecting $f(x_1)$ and $f(x_2)$. A function is convex when it's second derivative is non negative.

Jensen's Inequality: for f a convex function:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{2}$$

Data Processing Inequality: Suppose $X \to Y \to Z$ forms a Markov Chain. Then:

$$I(X;Y) \geq I(X;Z) \tag{3}$$

2 Chapter Three: AEP

2.1 Definitions

Markov's Inequality: deviation of a r.v. from some value

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{4}$$

Chebyshev's Inequality: deviation of r.v. from its mean:

$$\Pr(|X - \mu| \geq \varepsilon) \leq \frac{Var[X]}{\varepsilon^2} \tag{5}$$

Convergence in Probability: for a given sequence $\{A_i\}_{i=1}^{\infty}$:

$$\lim_{n\to\infty} \Pr(|A_i - L| > \varepsilon) = 0$$
$$\triangleq A_i \underset{\Pr}{\to} L$$

Weak Law of Large Numbers: if $\{X_i\}_{i=1}^{n}$ are iid random variables with mean μ and variance $\sigma^2 < \infty$, then the *sample mean* approaches the true mean as you get more samples:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{\Pr}{\to} \mu = \mathbb{E}[X] \tag{6}$$

2.2 AEP

AEP: Consider a sequence $\{X_i\}_{i=1}^{\infty}$ where each X_i is iid with pmf $p(x)$ and entropy $H(X)$. Then the *sample entropy* approaches the true entropy as you get more samples. Or:

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \underset{\Pr}{\to} H(X)$$
$$\lim_{n\to\infty} \Pr\left(\left|-\frac{1}{n}\log p(X_1, \ldots, X_n) - H(X)\right| > \varepsilon\right) = 0$$

2.3 Typical Set

Typical Set: contains all sequences with "sample entropy" $\approx H(X)$.

$$A_\varepsilon^{(n)} = \left\{x^n \in \mathcal{X}^n : \frac{1}{2^{n(H(x)+\varepsilon)}} \leq p(x^n) \leq \frac{1}{2^{n(H(x)-\varepsilon)}}\right\} \tag{7}$$

So the probability of the sequences is roughly the average probability sequence.

A few properties about $A_\varepsilon^{(n)}$:

(1) The first is that the "sample entropy" is close to the true entropy. That is, for each $x^n \in A_\varepsilon^{(n)}$:

$$H(X) - \varepsilon \leq -\frac{1}{n}\log p(x^n) \leq H(X) + \varepsilon \tag{8}$$

(2) Sampling a sequence from $\mathcal{X}^n$ has probability greater than $1-\varepsilon$ to be in the typical set:

$$\Pr\left(x^n \in A_\varepsilon^{(n)}\right) > 1-\varepsilon \tag{9}$$

Which follows from the AEP, basically. That is, since the sample entropy converges to $H(X)$ in probability, there must exist a n_ε such that:

$$\lim_{n\to\infty} \Pr(\underbrace{|-\frac{1}{n_\varepsilon}\log p(x^n) - H(X)|}_{①} > \varepsilon) < \delta \tag{10}$$

But note that the actual event, term ①, can be rewritten into $x^n \in A_\varepsilon^{(n)}$:

$$\begin{aligned}
-\varepsilon \leq -\frac{1}{n_\varepsilon}\log p(x^n) - H(X) &\leq \varepsilon \\
H(X) - \varepsilon \leq -\frac{1}{n_\varepsilon}\log p(x^n) &\leq \varepsilon + H(X) \\
-n_\varepsilon(H(X)-\varepsilon) \geq \log p(x^n) &\geq -n_\varepsilon(\varepsilon + H(X)) \\
2^{-n_\varepsilon(H(X)-\varepsilon)} \geq p(x^n) &\geq 2^{-n(\varepsilon+H(X))}
\end{aligned}$$

Which is exactly the condition for being in the typical set. Therefore, for n_ε, Equation 10 occurs, which implies that $x^n \in A_\varepsilon^{(n)}$. Therefore, the probability of being in the typical set goes to 1 for n sufficiently large.

(3,4) The last two properties give bounds on the size of the typical set:

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq A_\varepsilon^{(n)} \leq 2^{n(H(X)+\varepsilon)} \tag{11}$$

Where the left hand size (lower bound) is for n sufficiently large.

We know the total number of length n sequences is $|\mathcal{X}^n|$, but surely the typical set doesn't contain every length n sequence. Since we know that each $x^n \in A_\varepsilon^{(n)}$ has bounds on the probability of that sequence. We can leverage this to bound the size of $A_\varepsilon^{(n)}$:

$$\begin{aligned}
&\sum_{x^n\in\mathcal{X}^n} p(x^n) = 1 \\
\geq &\sum_{x^n\in A_\varepsilon^{(n)}} p(x^n) \\
\geq &\sum_{x^n\in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)}
\end{aligned}$$

But this last term doesn't depend on x^n, so:

$$\geq \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)} \tag{12}$$

From the first equality, we see:

$$\begin{aligned} |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)} &\leq 1 \\ \therefore \frac{|A_\varepsilon^{(n)}|}{2^{n(H(X)+\varepsilon)}} &\leq 1 \\ \therefore |A_\varepsilon^{(n)}| &\leq 2^{n(H(X)+\varepsilon)} \qquad \square \end{aligned}$$

Recall that $P(x^n \in A_\varepsilon^{(n)}) > 1 - \varepsilon$ for n large. We bound the LHS by:

$$P(x^n \in A_\varepsilon^{(n)}) \leq_M \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) = \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = |A_\varepsilon^{(n)} 2^{-n(H(X)-\varepsilon)} \tag{13}$$

Where $\leq_M$ follows since the right hand side gives the maximal probability of the set, since every element is maximally probable.

2.4 Codes

Code: Assigns a unique binary sequence to every sequence in $\mathcal{X}^n$.

Need:

- $n(H(X) + \varepsilon) + 2$ bits to make a code for each item in the typical set.
- $n \log |\mathcal{X}| + 1$ bits to make a code for each item *not* in the typical set.

Since there are $|A_\varepsilon^{(n)}| \leq 2^{-n(H(X)+\varepsilon)}$, if we just enumerate all items in the typical set, we need $n(H(X) + \varepsilon)$ bits to code each item. We then add 1 in case that's not an integer (could take ceil just easily), and add 1 so we prefix all typical sequences with a 0. Therefore we have a code that encodes all sequences in the typical set with $n(H(X) + \varepsilon) + 2$ bits. We get the same for the non typical sets, giving us a total code length of $n \log |\mathcal{X}| + 1$ (don't need +2 since guaranteed integer). If n is sufficiently large so that $\Pr(x^n \in A_\varepsilon^{(n)}) > 1 - \varepsilon$, the expected length of a code word in the typical set is:

$$\mathbb{E}\left[\frac{1}{n}\ell(X^n)\right] \leq H(X) + \varepsilon \tag{14}$$

On average, each element of the sequence takes about the entropy of the r.v. to encode. **Thus we can represent sequences X^n using around nH(X) bits on average**.

$$
\begin{aligned}
\mathbb{E}[\ell(X^n)] &= \sum_{x^n} p(x^n)\ell(x^n) \\
&= \left(\sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)(n(H(X)+\varepsilon)+2) \right) + \left(\sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)(n\log|\mathcal{X}|+1) \right) \\
&= \Pr(x^n \in A_\varepsilon{}^{(n)})(n(H(X)+\varepsilon)+2) + \Pr(x^n \notin A_\varepsilon^{(n)})(n\log|\mathcal{X}|+1) \\
&\leq (n(H(X)+\varepsilon)+2) + (n\log|\mathcal{X}|+1)
\end{aligned}
$$

2.5 Probable Set

Probable Set: what is the relationship between the typical set and the smallest such set that contains most the probability? Let $B_\delta^{(b)}$ be the smallest set such that:

$$\Pr(x^n \in B_\delta^{(b)}) \geq 1-\delta \tag{15}$$

Then we'll see that, for $\delta < \frac{1}{2}$ and $\delta' > 0$:

$$\frac{1}{n}\log\left|B_\delta^{(n)}\right| > H(X) - \delta' \tag{16}$$

Thus, $B_\delta^{(n)}$ must have at least $2^{nH(x)}$ elements, which is about the same size as $A_\varepsilon^{(n)}$.

3 Chapter Four: Entropy Rates, Stochastic Processes

3.1 Definitions

Stochastic Process: $\mathcal{S}$ is an indexed sequence of random variables, $\{X_k\}_{k=1}^{\infty}$

Stationary: A process $\mathcal{S}$ is stationary if the statistics don't change as you move in time:

$$\Pr(X_{k+1} = x_1, \ldots, X_{k+m} = x_m) = \Pr(X_1 = x_1, \ldots, X_m = x_m) \tag{17}$$

For all choices of m, k, and x.

Markov Process: $\mathcal{S}$ is a Markov Process (or Markov Chain) if:

$$\Pr(X_{k+1} = x_{k+1} \mid X_1, \ldots, X_k) = \Pr(X_{k+1} = x_{k+1} \mid X_k) \tag{18}$$

Time Invariance: property of a Markov Chain if $P(X_n \mid X_{n-1}) = P(X_2 \mid X_1)$.

Markov Chain can be characterized by a transition matrix, P:

$$P = \begin{bmatrix} P_{1,1} & P_{2,1} & \ldots & P_{n,1} \\ \vdots & \ddots & & \vdots \\ P_{1,n} & P_{1,n} & \ldots & P_{1,n} \end{bmatrix} \tag{19}$$

And a start state. Typically we'll ask for a start distribution. If the distribution after one transition is identical to the start distribution, then we say it's the **stationary distribution**:

$$\left[\mu_1, \mu_2, \ldots, \mu_n\right] = \left[\mu_1, \mu_2, \ldots, \mu_n\right] P \tag{20}$$

We solve for the stationary distribution using Eigenvalue decomposition with eigenvalue one, or just solving the system of equations.

That is, to solve for eigen values, we do $Av = \lambda v$, so $det(A - \lambda \mathcal{I})v = 0$, which gives the characteristic polynomial. Solve for the roots gives eigen vectors.

With the stationary distribution, we see $\mu = \mu P$, so μ is already known to be an eigenvector with eigenvalue 1. Thus, we just solve $det(P - \lambda I)\mu = 0$, for $\lambda = 1$.

3.2 Entropy Rates

Q: How does the entropy of a sequence $X_1, \ldots, X_n$ grow with n?

Entropy Rate: The per symbol entropy of the n random variables, when the limit exists:

$$H(\mathcal{S}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{21}$$

And a related quantity, the conditional entropy rate of the last random variable given the sequence:

$$H'(\mathcal{S}) = \lim_{n \to \infty} H(X_n \mid X_{n-1}, \ldots, X_1) \tag{22}$$

For a *stationary stochastic process*, $H(\mathcal{S}) = H'(\mathcal{S})$, and the limit exists. $H'(\mathcal{S})$ existing follows by conditioning reducing entropy and non-negativity of entropy, so the probability has to pile up. The second one, $H(\mathcal{S})$, follows by applying the chain rule, so we get a running average of conditional entropies. By the Cesaro Mean, a running average of things that converge to B also converges to B. Thus, $H(\mathcal{S})$ converges to $H'(\mathcal{S})$, so they're equal and the limit exists.

Entropy Rates. We have two definitions, $H(\mathcal{S})$ and $H'(\mathcal{S})$. Moreover, they're equivalent which is convenient for computing the entropy rate of a stationary Markov chain.

If $\mathcal{S}$ is a stationary Markov chain, then $H(\mathcal{S})$ is:

$$\begin{aligned} H(\mathcal{S}) = H'(\mathcal{S}) &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \\ &=_M \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\ &=_T \lim_{n\to\infty} H(X_2 \mid X_1) \\ &= H(X_2 \mid X_1) \end{aligned}$$

Where $=_M$ follows from the Markov property and $=_T$ follows by time invariance.

So the entropy rate of a stationary Markov chain is $H(X_2 \mid X_1)$. Let μ be the stationary distribution and P be the transition matrix. Then:

$$H(X_2 \mid X_1) = -\sum_{x_1\in\mathcal{X}} \sum_{x_2\in\mathcal{X}} p(x_1, x_2) \log p(x_2 \mid x_1) \tag{23}$$

Where by the chain rule of probability, $p(x_1, x_2) = p(x_2 \mid x_1)p(x_1)$. Note that the transition matrix P denotes the probability of going to state x_2 given state x_1, and μ denotes the probability of being in state x_1. So:

$$H(X_2 \mid X_1) = -\sum_{\mu_i} \sum_{P_{ij}} \mu_i P_{ij} \log P_{ij} \tag{24}$$

3.3 Thermodynamics

Relative entropy between two distributions decreases with time:

$$D(\mu_n \mid\mid \mu'_n) \geq D(\mu_{n+1} \mid\mid \mu'_{n+1}) \tag{25}$$

Argument follows from chain rule for entropy (or expanding and using total law of prob).

From this we also see that the relative entropy between any distribution and the stationary distribution decreases with time. Let $\mu'_n = \kappa$ be stationary, then $\mu'_{n+1} = \mu'_n$, so, applying the previous result:

$$D(\mu_n \mid\mid \kappa) \geq D(\mu_{n+1} \mid\mid \kappa) \tag{26}$$

3.4 Main Theorems

- Theorem 4.2.1: stationary stochastic process, limits of H and H' exist and are equal in the limit.

- 4.2.2: $H(X_n \mid X_{n-1} \dots X_1)$ is non-increasing in n and has limit H'.
- Cesaro Means
- Entropy Rate of Stationary Markov Chains
- Formula for the previous one
- Random Walks

4 Chapter Five: Data Compression

Definitions:

- A **Source Code** for a r.v. is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$, the set of finite-length strings from a size D alphabet. $C(x)$ is the codeword of x and $\ell(x)$ is the length of $C(x)$.
- A code word's **expected length** is: $L(C) = \sum_{x \in \mathcal{X}} p(x)\ell(x)$

4.1 Types of Codes

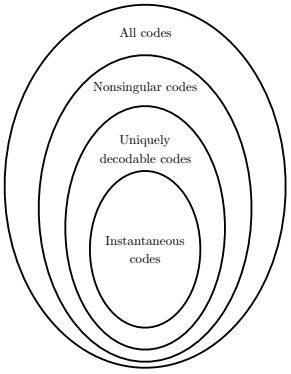
Three different types of **source** codes:

1. *Nonsingular*: every element of the alphabet of X maps to a different codeword:

$$x \neq x' \rightarrow C(x) \neq C(x') \tag{27}$$

2. *Uniquely Decodable*: Every sequence of coded strings decodes to exactly one message.
3. *Instantaneous/Prefix*: Can read each character as you read the code.

Where *Prefix* $\subset$ *Uniquely Decodable* $\subset$ *Nonsingular* $\subset$ *Codes*:



4.2 Minimizing Length

So what we really want are prefix/instantaneous codes (they have the nicest properties). Thus, our whole goal with source codes is to come up with the prefix coding scheme that yields the shortest possible expected length. Clearly we can't make every single codeword super short and still be a prefix code. The Kraft Inequality gives us a fundamental limitation on the length of codewords:

Theorem *(Kraft Inequality)*: Assume $\mathcal{C}$ is a prefix code. Then:

$$\sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1 \tag{28}$$

Theorem *(Converse Kraft Inequality)*: Given a set of lengths $\ell(x)$ that satisfy:

$$\sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1 \tag{29}$$

There exists a prefix code with these lengths.

Proof.

> We can think of a prefix code as a binary tree, where each branch represents choosing one of the D symbols for the next symbol of the code. Then a prefix code guarantees that each codeword has no children in the tree.
>
> Consider the length of the longest codeword ℓ_{max}. Now consider all codewords at this level of the tree.
>
> In the complete tree (so no children are pruned, they're just listed as a "descendent"), we have:
>
> $$\sum D^{\ell_{max} - \ell_i} \leq D^{\ell_{max}} \tag{30}$$

□

Converse Proof

Proof.

> Given lengths, $\ell_1 \ldots, \ell_k$ that satisfy the Kraft inequality, we can always come up with a prefix tree. □

We care about finding the prefix code with minimum expected length: that is, due to Kraft, we want to find a prefix code that satisfies the kraft inequality, that minimizes the expected code word length. So:

$$\min_{\ell}(L) = \min_{\ell} \left(\sum_i p_i \ell_i \right) \tag{31}$$

Over all integers $\ell_1 \ldots$ satisfying:

$$\sum_i D^{-\ell_i} \leq 1 \tag{32}$$

We solve this using **Lagrange Multipliers**:

$$\min_{\lambda} J = \min_{\lambda} \left(\sum p_i \ell_i + \lambda \left(\sum D^{-\ell_i} \right) \right) \tag{33}$$

Where we differentiate w.r.t. ℓ_i, to get:

$$\frac{\partial J}{\partial \ell_i} = p_i - \lambda D^{-\ell_i} log_e D \tag{34}$$

We set this equal to 0 to get:

$$D^{-\ell_i} = \frac{p_i}{\lambda \log_e D} \tag{35}$$

Now, we revisit our constraint from the Kraft Inequality:

$$\sum D^{-\ell_i} = 1$$
$$\therefore \sum \frac{p_i}{\lambda \log_e D} = \frac{1}{\lambda \log_e D} = 1$$
$$\therefore \lambda = \frac{1}{\log_e D}$$

So we conclude that $p_i = D^{-\ell_i}$.

Thus, the optimal code lengths are $\ell_i = \log_D \frac{1}{p_i}$. Later we'll force this to an integer with the ceiling operator.

Theorem: Expected length L of any instantaneous D-ary code for a r.v. X is lower bounded below by the entropy $H_D(X)$:

$$H_D(X) \leq L \leq H_D(X) + 1 \tag{36}$$

Proof of lower bound idea: Write out the difference $L - H_D(X)$ and turn the result into a relative entropy quantity plus a positive constant, by the information inequality we conclude $L - H_D(X) \geq 0$.

Proof of upper bound idea: Let each length $\ell_i = \lceil \log_D \frac{1}{p_i} \rceil$, so it's guaranteed to be between $\log_D \frac{1}{p_i}$ and $\log_D \frac{1}{p_i} + 1$. Then we multiplty both sides by p_i and sum over i to get the bounds.

So the entropy is the central limitation on Data Compression.

4.3 Huffman Codes

Definition 1 (Huffman Code): *The* **Huffman code** *is an algorithm for computing an optimal prefix code:*

INPUT*: a pmf, p, and an alphabet, $\mathcal{A}$.*
OUTPUT*: a code.*

There are two steps:

(1) Cluster

(2) Rerank

The cluster step takes a probability vector of m elements in order of mass: $\langle p_1, p_2, \ldots, p_m \rangle$ and computes an length $m-1$ vector, also in order, where the two smallest elements are merged: $\langle p_1, p_2, \ldots, p_{m-1} + p_m \rangle$.

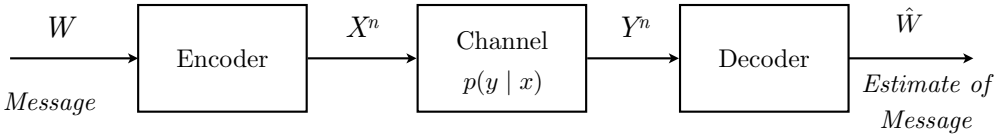
- Procedure for making them
- Examples
- They're optimal

4.4 Dyadic distributions

Generating distributions from fair coins, properties.

5 Chapter Seven: Channel Capacity

Summary: we find the maximum number of distinguishable signals for n uses of a communication channel.



Definition 2 (Channel Capacity): *We define the* **information channel capacity** *of a discrete memoryless channel (DMC) as:*

$$\max_{p(x)} I(X;Y) \tag{37}$$

Intuition: if we could control exactly how bits are sent, what's the most info we can send over the channel? How much shared info is there between the output and the input.

5.1 Example Channels

- Noiseless binary channel, Noisy channel with nonoverlapping outputs, Noise typewriter
- Binary Symmetric Channel with crossover probability p:

$$I(X;Y) = H(Y) - H(Y \mid X) = H(Y) - H(p) \leq 1 - H(p) \tag{38}$$

- Binary Erasure Channel with erasure probability α:

$$I(X;Y) = 1 - \alpha \tag{39}$$

5.2 Symmetric Channels

Definition 3 (Symmetric Channel): *A channel is said to be* **Symmetric** *if the rows of the channel transition matrix $p(y \mid x)$ are permutations of each other and the columns are permutations of each other.*

Definition 4 (Weakly Symmetric Channel): *A channel is said to be* **Weakly Symmetric** *if if every row of the transition matrix $p(\cdot \mid x)$ is a permutation of every other row and the sums $\sum_x p(y \mid x)$ are all equal.*

Theorem: For a weakly symmetric channel:

$$C = \log |\mathcal{Y}| - H(\text{row of transition matrix}) \tag{40}$$

Which is achieved by a uniform distribution on the inputs.

5.3 Properties of Channel Capacity

1. $C \geq 0$ since $I(X;Y) \geq 0$
2. $C \leq \min\{\log |\mathcal{X}|, \log |\mathcal{Y}|\}$, $C \leq \min\{H(X), H(Y)\}$.
3. $I(X;Y)$ is a continuous function of $p(x)$ and is concave.

5.4 Channel Coding Theorem

Definition 5 ((M,n) Code): *An* **(M,n) code** *for the channel $(\mathcal{X}, p(y \mid x), \mathcal{Y})$ consists of the following:*

1. *An index set $\{1,\dots,M\}$.*
2. *An encoding function $X^n : \{1,\dots,M\} \longmapsto \mathcal{X}^n$, resulting in codewords $x^n(1), x^n(2), \dots$. The set of codewords is called the code book.*
3. *A decoding function $g : \mathcal{Y}^n \longmapsto \{1,\dots,M\}$.*

5.4.1 Error and Rate

We have a several definitions relevant to error.

1. First: λ_i, which is the probability of error in sending message i over the channel:

$$\lambda_i = \Pr(g(Y^n) \neq i \mid X^n(i) = x^n(i)) = \sum_{y^n} p(y^n \mid x^n) \mathbb{1}\{y^n \neq g(x^n)\} \tag{41}$$

2. The maximal probability of error is just the maximal error term over all λ_i:

$$\lambda^{(n)} = \max_i \lambda_i \tag{42}$$

3. The average probability of error $P_e^{(n)}$ for an (M,n) code is:

$$P_e^{(n)} = \frac{1}{M} \sum_i \lambda_i \tag{43}$$

Definition 6 (Rate): *The* **rate**, *denoted R, of an (M, n) code is:*

$$R = \frac{\log M}{n} \qquad \textit{bits per transmission} \tag{44}$$

So the rate is, per bit sent over the channel, how much of the actual message does it actually capture?

A rate is *achievable* if there exists a sequence of $\left(\lceil 2^{nR} \rceil, n\right)$ codes such that the maximal probability of error goes to zero as n goes to infinity.

5.4.2 Jointly Typical Sets and Joint AEP

Definition 7 (Jointly Typical Set): *The* **jointly typical set** *for two r.v.'s is:*

$$A_\varepsilon^{(n)} \triangleq \{(x^n, y^n) : f_\varepsilon(x^n, y^n)\} \tag{45}$$

Where:

$$f_\varepsilon(x^n, y^n) = -\frac{1}{n} \log p(\bigcirc) - H(\bigcirc) < \varepsilon \tag{46}$$

Where $\bigcirc$ can be x^n, or can be y^n, or both at the same time (x^n, y^n). Where:

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \tag{47}$$

That is, it must satisfy all three conditions.

Joint AEP gives us the same properties as the AEP:

1. $\Pr((X^n, Y^n)) \in A_\varepsilon^{(n)} \to 1$ as $n \to \infty$
2. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$
3. Consider $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$: that is, the tilde vars are sampled independently but with the same marginals as X^n and Y^n. Then:

$$(1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)} \leq \Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\varepsilon)} \tag{48}$$

Takeaway from 3. is that, the probability that the independently sampled var-pairs are in the typical set is controlled by the mutual information.

Channel Coding Intuition: All rates below capacity C are achievable, and all rates above capacity are not.

Theorem *(Channel Coding Theorem)*: For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.

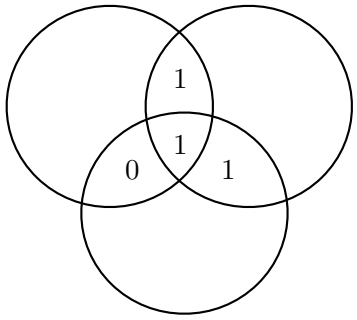
Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof idea:

- Pick a random codebook.
- Send messages as usual over the channel: so we receive y^n.
- Decode via joint typicality:
  - Search the codebook for the pair $(x^n(i), y^n) \in A_\varepsilon^{(n)}$. That is, find the input message that, when coded, probably led to the output.
  - We assume this is the message sent. We might error due to two things:
    1. We don't find such a pair.
    2. We find a pair but it's the wrong one.
  - Per the properties of the jointly typical set, both of these occur negligibly often.

5.5 Hamming Codes

The actual coding scheme used for the Channel Coding Theorem is highly impractical (it's random!). Hamming codes solve that. Best description is from the Venn Diagram:



Place the 4 information bits into the 4 central intersecting regions. To code, place 1s in each of the remaining regions so that each circle has an even number of bits. Then when you receive the message, reconstruct the venn diagrams and you can identify where bits may have been flipped.

Hamming code is the elements of the **null space** of the matrix denoting the possible messages. That is, each column is a possible message. We compute the null space of matrix h, which is the set of vectors such that $Hv = 0$.

Just solve: $Hv = 0$ and you're done.

5.6 Source Channel Coding Theorem

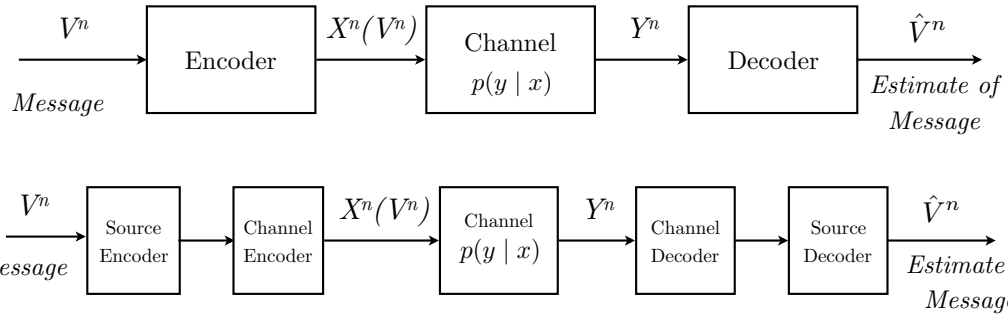
Here we combine the two central results:

1. Data compression: $R > H$
2. Data Transmission: $R < C$

Theorem *(Source Channel Coding Theorem)* Let $\mathcal{V} = V_1, V_2, \ldots, V_n$ be any stochastic process that satisfies the AEP, and $H(\mathcal{V}) < C$. Then there is a source-channel code with probability of error $\Pr(\widehat{V}^n \neq V^n) \to 0$.

Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, it's not possible to send the process over the channel with arbitrarily low probability of error.

Takeaway: The separation theorem says that the separate encoder can achieve the same rates as the joint encoder. That is, the following two are the same:



Proof idea:

- Since the stochastic process satisfies the AEP, it implies there exists a typical set.
- Index all sequences in the typical set.
- There are at most $2^{n(H(X)+\varepsilon)}$ elements in the typical set, so we need at most $n(H(X) + \varepsilon)$ bits to encode them.
- If $H(\mathcal{V}) + \varepsilon = R < C$, we can transmit the sequence with low probability of error:

$$\Pr(V^n \neq \widehat{V}^n) \leq \Pr(V^n \notin A_\varepsilon^{(n)}) + \Pr(g(Y^n) \neq V^n \mid V^n \in A_\varepsilon^{(n)}) \leq \varepsilon + \varepsilon \tag{49}$$

Converse combines Fano's Inequality and the Data Processing Inequality.

Fano's Inequality: For any estimator $\widehat{X}$ such that $X \to Y \to \widehat{X}$, with $P_e = \Pr(X \neq \widehat{X})$, we have:

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X \mid \widehat{X}) \geq H(X \mid Y) \tag{50}$$

Or:

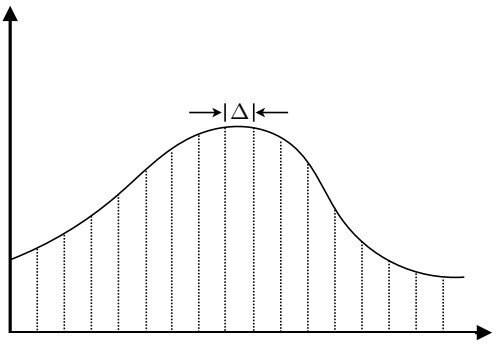
$$1 + P_e \log |\mathcal{X}| \geq H(X \mid Y) \tag{51}$$

6 Chapter Eight: Differential Entropy

Suppose X is a r.v. taking values in $\mathbb{R}$ with pdf $f(x)$, with support $\{x \mid f(x) > 0\}$. Then we get our usual definitions:

1. Differential Entropy: $h(X) = \int_{\mathcal{S}} f(x) \log f(x) dx$.
2. Joint Entropy: $h(X_1, X_2, \ldots, X_n) = -\int f(x^n) \log f(x^n) dx^n$.
3. Conditional Entropy: $h(X \mid Y) = -\int f(x, y) \log f(x \mid y) dx\ dy$.
4. KL-Divergence: $D(f \mid\mid g) = \int_{\mathcal{S}_f \cap \mathcal{S}_g} f(x) \log \frac{f(x)}{g(x)} dx$
5. Mutual Information: $I(X;Y) = \int f(x, y) \log \frac{f(x,y)}{f(x)f(y)} dx\ dy$

We can translate between Differential Entropy and Discrete Entropy. Consider quantizing $f(x)$ according to some fixed step size, Δ. That is, approximate the curve with blocks of width Δ.



Let $H(X_\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i$. Where p_i is going to be the value of those rectangles.

Consider a point x_k. Then along the interval, $[x_k, x_{k+\Delta}]$, let $p(x_k) = \int_{x_k}^{x_{k+\Delta}} f(x) dx$, where $H(X_\Delta) = -\sum_{-\infty}^{\infty} p(x_k) \log p(x_k)$.

Idea: We're taking rectangles and putting them over each interval so that the area of the rectangle is identical to the area of the curved piece of the function.

But $p(x_k) = p_k = f(x_k) \cdot \Delta$, since it just describes a box that approximates the pdf for that interval (width Δ and height $f(x_k)$). So:

$$H(X_\Delta) \approx h(x) - \log \Delta \tag{52}$$

In more detail, we have that:

$$\begin{aligned} H(X_\Delta) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\ &= -\sum_{-\infty}^{\infty} f(x_i)\Delta(\log f(x_i) + \log\Delta) \\ &= \underbrace{-\sum_{-\infty}^{\infty} f(x_i)\Delta \log f(x_i)}_{\text{goes to } h(f) \text{ as } \Delta \to 0} \underbrace{-\sum_{-\infty}^{\infty} f(x_i)\Delta \log \Delta}_{=\log\Delta} \end{aligned}$$

Therefore, we add the right term to get $H(X_\Delta) + \log\Delta = -\sum_{-\infty}^{\infty} f(x_i)\Delta \log f(x_i)$, which, as $\Delta \to 0$, becomes $h(X)$.

6.1 Examples

Now, some example continuous channels.

6.1.1 Uniform

Let X be a r.v. with uniform probability on the interval $[0, a]$. Then:

$$h(X) = -\int_0^a f(x) log f(x) dx = -\int_0^a \frac{1}{a} log \frac{1}{a} dx = \log a \tag{53}$$

6.1.2 Gaussian

Let X be a r.v. with a Gaussian density function:

$$f(x) \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{-x^2}{2\sigma^2} \tag{54}$$

Then it's entropy is:

$$h(X) = -\int_{-\infty}^{\infty} \phi \ln \phi = \frac{1}{2} \ln 2\pi e\sigma^2 \tag{55}$$

6.2 Mutual Info of Multivariate Normal

Suppose $K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$. Then:

$$I(X;Y) = h(X) + h(Y) - h(X,Y) \tag{56}$$

And we can compute $h(X)$ and $h(Y)$ via the entropy of a normal distribution: $\frac{1}{2}\log\left[2\pi e\sigma^2\right]$, and we can compute $h(X,Y)$ as the entropy of a multivariate normal: $\frac{1}{2}\log\left[2\pi e \det(K)\right]$, and we're done.

6.3 Multivariate Normal

$$h(X_1, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log \left[2\pi e \det(K)\right] \tag{57}$$

6.4 Identities

Key: In general, $h(X) + n$ is the number of bits on the average required to describe X to n-bit accuracy.

- Venn Diagram: $I(X;Y) = h(X) - h(Y \mid X) = h(Y) - h(X \mid Y) = h(X) + h(Y) - h(X,Y)$
- Information Inequality: $D(f \mid\mid g) \geq 0$. Consequently:
  - $I(X;Y) \geq 0$, equality iff independent.
  - $h(X) \geq h(X \mid Y)$, equality iff independent.
  - $h(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i)$, equality iff independent.
- Hadamard's Inequality:

$$\det(K) \leq \prod_{i=1}^{n} K_{i,i} \tag{58}$$

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$

6.5 AEP For Continuous R.Vs

Let $X_1, \ldots, X_n$ be a sequence of r.v.s drawn i.i.d. from density $f(x)$. Then:

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) \to h(X) \tag{59}$$

Typical set is the same, but the set-cardinality is translated into the Volume of the continuous set.

6.6 Maximum Entropy Bound

If a pdf has variance N, then the entropy of the pdf is upper bounded by:

$$h(f) \leq h(\mathcal{N}(0, N)) = \frac{1}{2} \log \left[2\pi e N\right] \tag{60}$$

Proof idea:

- Consider the relative entropy of a pdf with variance/covariance K and a normal with variance/covariance K: ϕ_K. This is: $D(f||\phi_K)$.
- Expanding, we get:

$$0 \leq D(f||\phi_K) = \int f(x) \log \frac{f}{\phi_K} dx = \int f \log f - \int f \log \phi_K = -h(f) + h(\phi_k) = h(\phi_k) - h(f) \tag{61}$$

  And we know the last piece is ≥ 0, so we're done.

7 Chapter Nine: Gaussian Channel

X_i is a r.v. with a continuous alphabet $\mathcal{X}$, we have a time discrete channel with noise $Y_i = X_i + Z_i$, where the noise $Z_i \sim \mathcal{N}(0, N)$. The noise is assumed to be independent of the signal.

If there's no constraint on the input, then the capacity could be infinite since X can take any real value, so we can just spread the input values arbitrarily far apart subject to whatever noise is present in the channel. To avoid this (which is clearly unrealistic) we impose an input power constraint.

Power Constraint:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P \tag{62}$$

Capacity is the same, but now subject to a power constraint:

$$C = \max_{p(x):\mathbb{E}[x]\leq P} I(X;Y) \leq \frac{1}{2}\pi e\left(1 + \frac{P}{N}\right) \tag{63}$$

Proof idea is just to write it out: The noise is just Z, so $I(X;Y) = h(Y) - h(Y \mid X) = h(Y) - h(Z)$, where $h(Z) = \frac{1}{2}\log[2\pi eN]$. Plug and chug.

Note that:

$$\mathbb{E}[Y^2] = \mathbb{E}[(X+Z)^2] = \mathbb{E}[X^2 + XZ + Z^2] = \mathbb{E}[X^2] + \mathbb{E}[XZ] + \mathbb{E}[Z^2] \tag{64}$$

We know $\mathbb{E}[Z] = 0$,

7.1 Codes

We can make codes in the same way, only now the encoding function produces codewords such that: $\sum_{i=1}^{n} x_i(w)^2 \leq nP$.

A rate is achievable there exists a code that satisfies the power constraint and the usual notion of achievability is obtained.

Channel Coding Theorem: We get the channel coding theorem again for Gaussian channels. That is, any Rate $R < C$ is achievable, and the converse, that any rate $R \geq C$ is not achievable.

7.2 Band Limited Channels

Consider the frequency domain, with ω ranging from $-\infty$ to ∞, corresponding to different frequencies. Consider a continuous function of time $f(t)$ that spits out different frequencies.

Definition 8 (Bandlimited): *We say $f(t)$ is* **bandlimited** *to W if $F(\omega) = 0$ for $|\omega| > W$.*

So if there is some value for which, outside that interval, there are no frequencies!

7.3 Shannon-Nyquist Theorem

Idea: if f is bandlimited to W (so it only has frequencies in the interval $[-2\pi W, 2\pi W]$), from discrete samples we can reconstruct the full continuous function. That is, we can reconstruct the full signal from samples taken at every $\frac{1}{2W}$ seconds. So the full function $f(t)$ is determined by $f\left(\frac{n}{2W}\right)$, for $n \in \mathbb{Z}$.

$$f(n/2W) = 1/2\pi \int_{-2\pi W}^{2\pi W} F(\omega) \exp(i\omega \frac{n}{2W}) dw \tag{65}$$

.....

□

References

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.