EE 276

Taught by Tsachy Weissman
Notes by Chris Fifty

Winter 2024

Contents

1 January 11

1.1 Administrative

Homework 50%, Midterm 20%, Final 30%, participation 5%.

Lecture 2 winter 2020 lecture videos.

1.2 Notation and Review of Probability

Denote probability space $(\Omega, \mathcal{F}, P)$ as the sample space, event space, and probability measure.

1. Ω is the set of all outcomes for a trial. Rolling a die has 6 outcomes: $\{1, 2, 3, 4, 5, 6\}$.
2. $\mathcal{F}$ is a subset of the set of all possible outcomes. An event is a set of outcomes in the sample space, and the event space is a set of these events. For example: we could define one event to be even rolls $\{2, 4, 6\}$, another to be rolling a 1: $\{1\}$, and our event space could consist of both events. A simple event space is the set of all subsets of the sample space.
3. P is a probability measure: $\mathcal{F} \to [0, \infty)$.

Definition 1 (Random Variable)**.** *A mapping $X : \Omega \to E$ between two measurable spaces $(\Omega, \mathcal{F})$ and (E, ϵ) where ϵ is a $\sigma-$algebra on E.*

Denote X as a random variable or object over $\mathcal{X}$ sample space: $\mathcal{X}$ is the alphabet or values X can adopt. x is a particular value that a random variable X can take. For a discrete random variable X: $P_X^{(x)} = P(X = x) = p(x)$ probability that RV X adopts value x.

Theorem 1 (Jensen's Inequality)**.** *$\forall$ random variable X and any convex function f:*

$$E[f(X)] \geq f(E[X])$$

Lemma 1 (Degenerate Jensen's)**.** *If f is strictly convex, then $E[f(X)] = f(E[X]) \iff X$ is deterministic (takes on a single value with probability 1).*

1.3 Measures of Information

Suppose U is a random variable (or object) taking on values in a discrete and finite alphabet $\mathcal{U} = \{1, 2, ..., 3\}$. Can we associate a notion of surprise that U takes on a specific value?

How surprising would it be if U takes on the value u? As the probability of u becomes smaller, will it be more or less surprising to me that the realized value is u? Want to associate a measure of surprise to this event occurring.

Definition 2 (Surprise)**.** *We say $S(u) = \log \frac{1}{p(u)}$: The surprise function or the self-information, or log-loss. We assume base-2 log unless otherwise noted. Intuitively, the lower the probability that $U = u$, the higher $S(u)$, or our "surprise" that U takes on this value.*

Definition 3 (Entropy of a Random Variable)**.** *The entropy of a random variable U is the expected value of the surprise of U:*

$$\begin{aligned} H(U) &:= E[S(U)] \\ &= \sum_{u \in U} p(u)S(u) \\ &= \sum_{u \in U} p(U) \log \frac{1}{p(U)} \end{aligned}$$

Proposition 1. *For a random variable U that can take on r different values, $H(U) = \log(r)$ when U is distributed according to a uniform distribution: $p(u) = \frac{1}{r}$ for all $u \in \Omega$.*

Proof.

$$\begin{aligned} H(U) &= E[\log \frac{1}{p(U)}] \\ &= \sum_{u \in U} \frac{1}{r} \log r \\ &= \log r \end{aligned}$$

□

Recall that Jensen's inequality becomes an equality for a strictly concave function $\iff$ the random variable is deterministic. For the random variable $S(U)$—a function of a random variable is itself a RV—$S(U = u)$ is deterministic (i.e. adopts a single value) exactly $p(U = u)$ is the same for all $u \in \Omega$.

Proposition 2. *$H(U) = 0 \iff$ the random variable U is deterministic.*

Unlike the previous proposition where $p(U)$ was deterministic, now U is deterministic (so $p(U) = 1$).

Proof. Suppose U takes on a single value. Then $S(U) = 0$ and $E[S(U)] = 0$, so $H(U) = 0$.

Now suppose $H(U) = 0$. Then $E[S(U)] = 0 \iff S(U) = 0$. Because $S(U) \geq 0$, $S(U) = 0$ with probability 1 $\iff P(U) = 1$ deterministically $\iff U$ is deterministic. □

Consider a different pmf q. Let's say U is distributed according to pmf p; however, we think U is distributed according to q. The expected surprise that U is governed by q is:

Definition 4 (Cross-Entropy)**.**

$$H_q(U) := \mathbb{E}[\log \frac{1}{q(u)}]$$

Intuitively, the cross-entropy is the expected surprise if a non-optimal distribution is used to model $p(u)$. Accordingly, there is more "surprise" (or entropy) than if the correct distribution p was used.

Proposition 3. *$H(U) \leq H_q(U)$: the expected surprise from U (i.e. low probability events are down-weighted) is less than the surprise from U given we think the distribution comes from q. The equality holds $\iff p = q$ (i.e. U is distributed according to p, q).*

Definition 5 (KL Divergence or Relative Entropy)**.** *The relative entropy (or KL divergence) between two pmfs p, q is:*

$$\begin{aligned} D(p||q) &:= H_q(U) - H(U) \\ &= \mathbb{E}[\log \frac{1}{q(U)} - \log \frac{1}{p(U)}] \\ &= \sum_{u=1}^{r} p(u) \log \frac{p(u)}{q(u)} \end{aligned}$$

where both probability mass functions need to be over the same alphabet.

Remark 1.

$$H(U) = H_q(U) \iff D(p||q) = 0 \iff p = q$$

Specifically, the KL Divergence between two distributions (and therefore the difference between cross-entropy and entropy) is 0 if and only if distributions p and q are identically distributed. Accordingly, 0 is a lower bound on the KL-divergence. This measure is not upper bounded.

Definition 6 (Joint Entropy)**.** *For a pair of random variables (U, V), we define the joint entropy as:*

$$\begin{aligned}
H(U,V) &:= H((U,V)) \\
&= E[\log \frac{1}{p(u,v)}] \\
&= \sum_{u,v} p(u,v) \frac{1}{\log p(u,v)} \\
&= \sum_{u \in U} p(u) \sum_{v \in V} p(v|u) \log \frac{1}{p(u,v)}
\end{aligned}$$

as the entropy of the pair of random objects occurring together.

Definition 7 (Conditional Entropy)**.** *We can also define the conditional entropy as the entropy of a random variable U given the value of another random variable V is known:*

$$\begin{aligned}
H(U|V) &:= \mathbb{E}[\log \frac{1}{p(u|v)}] \\
&= \sum_{u,v} p(u,v) \log(\frac{1}{p(u|v)}) \\
&= \sum_{v} \sum_{u} p(v) p(u|v) \log \frac{1}{p(u|v)} \\
&= \sum_{v} p(v) \sum_{u} p(u|v) \log \frac{1}{p(u|v)}
\end{aligned}$$

The conditional entropy is simply the entropy of a random variable U with pmf distributed according to $p(u|v)$: $\sum_{u \in U} p(u|v) \log \frac{1}{p(u|v)} = H(U|V = v)$. In other words, the conditional entropy is when we average $H(U|V = v)$ over all possible values of V weighted by their relative probabilities of occurring (i.e. $p(v)$).

Property 1.3.1 (Chain Rule For Entropy)**.**

$$H(U,V) = H(U) + H(V|U)$$

Unlike the chain rule for probabilities that is multiplicative, the chain rule for entropy is additive.

Proof.

$$\begin{aligned}
H(U,V) &= \mathbb{E}[\log \frac{1}{P(U,V)}] \\
&= \mathbb{E}[\log \frac{1}{P(U)P(V|U)}] \\
&= \mathbb{E}[\log \frac{1}{P(U)}] + \mathbb{E}[\log \frac{1}{P(U|V}] \\
&= H(U) + H(V|U)
\end{aligned}$$

□

The log probability results in Entropy being additive rater than multiplicative.

Property 1.3.2. *$H(U|V) = H(U) \iff U, V$ are independent. Otherwise, $H(U) \geq H(U|V)$.*

Proof.

$$\begin{aligned}
H(U) - H(U|V) &= \mathbb{E}[\log \frac{1}{p(U)} - \log \frac{1}{p(U|V)}] \\
&= \mathbb{E}[\log \frac{p(U|V)}{p(U)}] \\
&= \mathbb{E}[\log \frac{p(U|V)}{p(U)} \frac{p(V)}{p(V)}] \\
&= \mathbb{E}[\log \frac{P(U,V)}{P(U)P(V)}] \\
&= \sum_{u,v} p(u,v) \log \frac{p(u,v)}{p(u)p(v)} \\
&= D(P_{(U,V)}||P_U \times P_V)
\end{aligned}$$

Where P_U is the probability mass function of U and $P_U \times P_V$ denotes the multiplication of their pmfs. So $P_{(u,v)} = P_u \times P_v \iff U$ and V are independent. This also supports our notion as the relative entropy (i.e. KL divergence) as a measure of distance between the pmfs: $P_{(U,V)}$ and $P_U \times P_V$. □

2 January 16

2.1 Administrative

1. Problem session: Packard 202 from 4-5:30 PM.

2.2 Lecture

Recap: entropy of a joint random variable $H(U)$, joint entropy of two random variables $H(U,V)$, conditional entropy $H(U|V)$. We've shown that conditioning reduces entropy: $H(U|V) \leq H(U)$ and moreover, $H(U) - H(U|V) = D(P_{U,V}||P_U \times P_V)$ where D is the relative entropy between the joint distribution and the distribution as if U, V were independent.

Definition 8 (Mutual Information)**.** *The mutual information between U and V is*

$$\begin{aligned}
I(U;V) &:= D(P_{U,V}||P_U \times P_V) \\
&= H(U) - H(U|V) \\
&= H(U) + H(V) - H(V) - H(U|V) \\
&= H(U) + H(V) - [H(U|V) + H(V)] \\
&= H(U) + H(V) - H(U,V) \\
&= H(U) + H(V) - [H(U) + H(V|U)] \\
&= H(V) - H(V|U) \\
&= I(V;U)
\end{aligned}$$

Where $H(U|V) + H(V) = H(U,V)$ via the chain rule.

Theorem 2 (Law of Large Numbers)**.** *For independent, identically distributed random variables (i.i.d), $\frac{1}{n}\sum_{i=1}^{n} X_i$ approaches the expected value $\mathbb{E}[X^n]$ as $n \to \infty$. Therefore,*

$$\begin{aligned}
\mathbb{E}[S(X)] &= \frac{1}{n} S(X) \\
&= \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{X_i} \\
&= \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{X_i} \\
&= \mathbb{E}[\log \frac{1}{X}] \\
&= H(X)
\end{aligned}$$

Asymptotic Equipartition Property (AEP): i.i.d. source $U_1, U_2, ...$ i.i.d. $\sim U$ for outcome space $\Omega = \{1, 2, .., r\}$. Denote $n-tuple$ $u^n = (u_1, u_2, ..., u_n)$ where $u_i \in \Omega$. $u^n \in \mathcal{F}$ the event space: a subset of sets of the outcome space.

Note: $p(u^n) = P(U^n = u^n) = \prod_{i=1}^n p(u_i)$ [as we assume each u_i is i.i.d].
Moreover $\log p(u^n) = \sum_{i=1}^n \log p(u_i)$.

Definition 9 (Typical). *u^n is* ***typical*** *if $p(u^n) \approx 2^{-nH(U)}$. More precisely, u^n is $\epsilon-typical$ if*

$$
\begin{aligned}
2^{-nH(U)+\epsilon} &\leq p(u^n) \leq 2^{-nH(U)-\epsilon} \\
\iff n[H(U) - \epsilon] &\leq -\log p(u^n) \leq n[H(U) + \epsilon] \\
\iff [H(U) - \epsilon] &\leq -\frac{1}{n} \log p(u^n) \leq [H(U) + \epsilon]
\end{aligned}
$$

Denote the set of all typical vectors by $A_\epsilon^{(n)}(U)$.

Theorem 3. *$\forall \epsilon > 0$, $P(U^n \in A_\epsilon^n(U))$ (i.e. the probability that the realized sequence u^n from RV U^n is actually typical)*

$$
\begin{aligned}
P(U^n \in A_\epsilon^n(U)) = P(|\frac{1}{n} \log \frac{1}{P(U^n)} - H(U)| \leq \epsilon) \\
\lim_{n \to \infty} P(U^n \in A_\epsilon^n(U)) = 1
\end{aligned}
$$

Proof.

$$
P(|\frac{1}{n} \log \frac{1}{P(U^n)} - H(U)| \leq \epsilon) = P(|\frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(U_1)} - H(U)| \leq \epsilon)
$$

where

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(U_i)} &\to \frac{1}{n} \sum_{i=1}^n S(U_i) \\
&\to \mathbb{E}[S(U)]
\end{aligned}
$$

where we used the law of large numbers to show the mean of the sample surprises for each RV U_i converges to the expected surprise of the underlying RV $U \sim p$ as $n \to \infty$. Then we have

$$
\begin{aligned}
P(|H(U) - H(U)| \leq \epsilon) \\
P(0 \leq \epsilon)
\end{aligned}
$$

□

We should expect that $|A_\epsilon^{(n)}(U)|$ (i.e. the order or size) $\approx 2^{nH}$.

Theorem 4. *For any $\epsilon > 0$ and n sufficiently large*

$$
(1 - \epsilon) 2^{n[H(U) - \epsilon]} \leq |A_\epsilon^{(n)}(U)| \leq 2^{n[H(U)+\epsilon]}
$$

Proof of upper bound.

$$
\begin{aligned}
1 &\geq P(U^n \in A_\epsilon^{(n)}(U)) \\
&= \sum_{u^n \in A_\epsilon^{(n)}} p(u^n) \geq \sum_{u^n \in A_\epsilon^{(n))}} 2^{-n[H+\epsilon]} \\
&= |A_\epsilon^{(n)}| \cdot 2^{-n[H+\epsilon]}
\end{aligned}
$$

□

Let $\mathcal{F} = \mathcal{U}^n$ be the event space over $n - tuples$ over the sample space. Then $\exists A_\epsilon^{(n)}$ the typical set $\subseteq \mathcal{F}$ as $P(U^n \in A_\epsilon^{(n)}) \approx 1$ as $n \to \infty$ as $|A_\epsilon^{(n)} \approx 2^{nH}$ and $\frac{|A_\epsilon^{(n)}}{|\mathcal{F}|} \approx \frac{2^{nH}}{r^n} = \frac{2^{nH}}{2^{n \log r}} = 2^{-n[\log r - H]}$, so the size of this set is exponentially small relative to all source sequences. Most individual sequences are outside this typical set; however somewhat paradoxically, if you look probabilistically, the sequence most likely to be realized is within the typical set. This is exactly what we call the AEP: when n is large, we effectivelly have a uniform distribution across the typical set, even though this set is relatively small compared to the event space. All of the probability mass is centered in the typical set.

Might there be a subset of the typical set (an even smaller set) that the probability mass in the event space centers around? Is there a set $|B_n| \leq 2^{N[H-\delta]} = |A_\epsilon^{(n)}|$? By theorem 1, the probability of $\mathcal{F} - A_\epsilon^{(n)}$ is less than $(1-\epsilon)2^{n[H-\epsilon]}$, so the part of B_n not in the typical set is less than this value. And the size of the intersection with the typical set is small since probability mass is uniformly distributed across the typical space (by definition).

Theorem 5. *For any $\alpha > 0$, and any $\{B_n\}_{n \geq 1}$, $B_n \subseteq \mathcal{F}$ such that $|B| \leq 2^{n[H(U)-\alpha]}$, we have $P(U^n \in B_n) \to 0$ as $n \to \infty$.*

Example 1. *$U \sim Ber(q)$ $0 < q < 1$, $q \neq \frac{1}{2}$. $P(U^n) = p^{N_1(U^n)} \cdot (1-q)^{N_0(U^n)}$ with N_1 as the number of 1s and N_0 as the number of 0s in u^n.*

$$\begin{aligned}\frac{1}{n} \log \frac{1}{P(U^n)} &= -\log p^{N_1(U^n)} \cdot (1-q)^{N_0(U^n)} \\ &= \frac{N_1(U^n)}{n} \log q + \frac{N_0(U^n)}{n} \log(1-q)\end{aligned}$$

By the law of large numbers $N_1 \to qn$, so

$$\begin{aligned}\frac{N_1(U^n)}{n} \log q + \frac{N_0(U^n)}{n} \log(1-q) &\approx q \log \frac{1}{q} + (1-q) \log \frac{1}{q} \\ &= H(U)\end{aligned}$$

Think about what happens when q = $\frac{1}{2}$: all sequences are typical.

3 January 18

Recap AEP: $U_1, U_2, ...$ are i.i.d. drawn from some distribution U. We're looking at an n-tuple $(U_1, U_2, ..., U_n)$ representing all possible source sequences. If probability space has size r, then event space has total size r^n. We've established the existence of a much smaller subset, the **typical set**, that has size exonentially smaller than the event space: 2^{nH}. Yet, with overwhealming probability, a source sequence sampled from U will come from the typical space. $P(U \in T) \approx 1$: the probability that U is in the typical space approaches 1 as $n \to \infty$. Moreover, any set of size 2^{nH} that is not the typical set has vanishingly probability of occurrence.

This has massive implications for compression. And the entropy is a bound as the typical set has size 2^{nH}.

Implications for compression: if your source sequence only comes from the typical set, would need nH bits to represent it since the size of the typical set is 2^{nH}. Need $\log |A_\epsilon^{(n)}| \approx nH$ bits to represent the typical $U^n s$.

For the entire event space, would need $\log r^n = n \log r$ to represent any sequence in the event space. This is much bigger!

Lossless compressor: one bit to tell you if a sequence is typical or not. If typical, invest nH bits. Otherwise, invest $n \log r$ bits.

Therefore, $\forall \epsilon > 0$ $\exists n$ and a compressor that will require no more than $\leq H + \epsilon$ bits per source symbol on average. On the other hand, by Theorem 3, we cannot achieve lossless compression for any scheme that uses less than $H - \delta$ bits per source symbol (only spend $n(H-\delta)$ bits to represent $2^{n(H-\delta)}$ possibilities which is less than the size of our typical set.

Take home message: Entropy is the fundamental limit on compression.

Variable Length Lossless Compression: Example: $\Omega = \{a, b, c, d\}$. $p(a) = \frac{1}{2}$, $p(b) = \frac{1}{4}$, $p(c) = \frac{1}{8}$, $p(d) = \frac{1}{8}$.

What is the optimal binary representation for this probability space?

Codewords: $c(a) = 0$, $c(b) = 10$, $c(c) = 110$, $c(d) = 111$. so the length of the codeword $\bar{l} = \mathbb{E}[l(U)] = 1.75$. Note that here: $l(u) = \log \frac{1}{p(u)}$ the length of a given codewod is equal to $\frac{1}{p(u)}$. $l(a) = \frac{1}{2} = 2^{-1}$. So $\mathbb{E}[l(U)] = E \log \frac{1}{P(U)} = H(U)$. This code actually achieves the entropy (so it is optimal).

Definition 10 (Uniquely Decodable)**.** *A code is uniquely decodable (UD) if every sequence of source symbols is mapped to a distinct binary representation.*

i.e. we could have $a \to 1$, $b \to 0$ and $c \to 10$ is not uniquely decodable.

Definition 11 (Prefix Code)**.** *A **prefix code** is one where no code word is the prefix of another. Prefix codes are uniquely decodable.*

Consdier the converse: do you need to be a prefix code to be uniquely decodable?
HW Exercise: Consider the following code: $\Omega = \{a, b, c, d\}$. $c(a) = 10$, $c(b) = 00$, $c(c) = 11$ $c(d) = 110$. Is it a prefix code? No – $c(c)$ is the prefix of $c(d)$. Is it uniquely decodable? It is! Will show this!

Dyadic source: probabilities are powers of 2^{-i} where $i \in \mathbb{Z}$.

Prefix codes for Dyadic distributions

Definition 12. *A source is dyadic if* $\forall u \in \Omega$, $\log \frac{1}{p(u)}$ *is an integer.*

Lemma 2. *Assume that* U *is dyadic with* $|\Omega| \geq 2$*. The number of symbols* $u \in \Omega$ *with* $p(u) = p_{\min} := \min_{u \in \Omega} p(u)$ *is even. The number of symbols that have the lowest probability is even: there will be 2, 4, 6, ... etc. symbols with the lowest probability.*

Proof. See course notes: intuitively, $\sum_i p_i = 1$ and each p_i is 2^{-i}. □

Prefix code construction (for dyadic source): Choose 2 symbols with $p(u) = p_{min}$ and merge them into one source symbol that has the sum of their probabilities. For our earlier example, this is equivalent to merging c, d into a single source symbol that has probability $\frac{1}{4}$. We now have a new source with this combined symbol.

This new source is also dyadic. We are going to repeat until we're left with a single symbol. This creates a binary tree, and then we can assign the leaves to the codewords. The number of merges is k because this is $\log 2^k$: the binary tree has k splits. Function of a dyadic sequence terminating when there are two elements with the same probability.

Note with this construction $l(u) = k = \log \frac{1}{p(u)}$ achieves the entropy precisely and tightly with $l = H(U)$.

Shannon Code:
For a general source, let $n_u = \log\lceil \frac{1}{p(u)} \rceil$ $\forall u \in \Omega$. Note that $\sum_{u\in\Omega} 2^{-n_u} = \sum_{u\in\Omega} 2^{-\lceil \log \frac{1}{p(u)} \rceil} \leq \sum_{u\in\Omega} 2^{-\log \frac{1}{p(u)}} = \sum_{u\in\Omega} p(u) = 1$.

You can think of Shannon Codes: for a new source distribution, we change the probabilities from $p(u)$ to 2^{-n_u}. Now have a dyadic source whose probabilities are 2^{-n_u}.

So can consider dyadic source U^* (a new random variable) with alphabet $\Omega^* \supseteq \Omega$ and $p^*(u) = 2^{-n_u}$ $\forall u \in \Omega$. We're basically translating the source symbols from our original alphabet Ω to have different probabilities in Ω^*. The probabilities for the symbols in the original alphabet are now dyadic!

This is exactly what the Shannon code is: the optimal prefix code for U^*. Denote length function $l_{Shannon}(u) = n_u = \lceil \log \frac{1}{p(u)} \rceil$ $\forall u \in \Omega$. Moreover, $\bar{l}_{Shannon} = \mathbb{E}[l_{Shannon}(U)] \leq \mathbb{E}[\log \frac{1}{p(U)} + 1] = H(U) + 1$. The shannon code will get us to within one bit of the entropy.

4 January 23

4.1 Recap: Prefix Codes and Shannon Codes

Prefix codes and how to construct prefix codes for Dyadic distributions: $p(u) = 2^{-n(u)}$ where n is an integer. Shannon codes: given a source $p(u)$, $n^*(u) = \lceil 2y \frac{1}{p(u)} \rceil$ where $p^*(u) = 2^{-n^*(u)}$.

E.g. Shannon code $\mathcal{U} = \{a, b\}$ $p(a) = 0.99$, so $n^*(u) = \lceil \log \frac{1}{0.99} \rceil$ and $p(b) = 0.01$ so $n^*(u) = 7$. So $p^*(u)$ is a dyadic distribution with $p^*(a) = \frac{1}{2}$ $p^*(b) = \frac{1}{128}$ and we create excess codewords to fill in the probability gaps: $p^*(c) = \frac{1}{128}$ ... $p^*(h) = \frac{1}{4}$. so $a \to 0$ and $b \to 1111110$. However; the shannon code neglected the simpliest code we could have chosen: set $a \to 0$ and $b \to 1$.

Last time we said $\mathbb{E}[L(U)] \leq H(U)+1$. However, for any uniquely decodable (UD) code, we must have $\mathbb{E}[L(X)] \geq H(X)$. Huffman code previously the best prefix code.

Note: Arithematic coding is typically taught in this course, but is skipped this year. In lecture video 6 from EE274.

Theorem 6 (Kraft-Mcmullan Inequality)**.** *The integer valued function $l(U)$ is the length function of a UD code $\iff \sum_{u \in U} 2^{-l(U)} \leq 1$. For every symbol, just take the size of the code . For code $C(u)$, $l(U) := |C(U)|$.*

For example, $c(a) = 110$ has $L(U = a) = 3$ and $2^{-3} = \frac{1}{8}$.

This is a a powerful inequality – it allows us to determine if a specific code is uniquely decodable.

Proof. If $l(U)$ satisfies the inequality, then define a dyadic distribution: $p^*(u) = 2^{-l(u)}$, and apply Shannon's code. Conversely, given a length function $l(U)$ of a uniquely decodable code. Define $l_{max} = \max_U l(U)$. Consider $(\sum_{u \in U} 2^{-l(U)})^k$ where $k > 0$ is an positive integer.

$$\begin{aligned}
(\sum_{u \in U} 2^{-l(U)})^k &= \sum_{U_1 \in U} 2^{-l(U_1)} \cdots \sum_{U_2 \in U} 2^{-l(U_k} \\
&= \sum_{U_1, \ldots, U_k} 2^{-\sum_{i=1}^{k} l(U_i)} \\
&= \sum_{u^K \in U^k} 2^{-l(U^K)} \\
&= \sum_{i=1}^{K \cdot l_{\max}} 2^{-i} |\{u_K | l(U_k) = i\}| \\
&\leq K \cdot l_{\max}
\end{aligned}$$

by unique decodability $|\{u_K | l(U_k) = i\}| \leq 2^i$. So we have $(\sum_{u \in U} 2^{-l(u)}) \leq (K l_{max})^{\frac{1}{K}} \to 1$ as $K \to \infty$. □

Conclusion: For any uniquely decodable code:

$$\mathbb{E}[l(U)] = \bar{l} \geq H(U)$$

for any $P_U(u)$.

Proof.

$$\begin{aligned}
\bar{l} - H(U) &= \sum_u p(u) \cdot l(u) + \sum_u p(u) \log(p(u)) \\
&= -\sum_u p(u) \log(2^{-l(u)}) + \sum_u p(u) \log p(u) \\
Z &= \sum_u 2^{-l(u)} \leq 1 \\
&= -\sum_u p(u) \log(\frac{2^{-l(u)}}{Z}) + \sum_u p(u) \log p(u) - \sum_u p(u) \log(Z) \\
q(u) &= \frac{2^{-l(u)}}{Z} \\
&= D(p||q) - \sum_u p(u) \log(Z) \\
&= D(p||q) - \log(Z) \\
&\geq 0
\end{aligned}$$

□

Remarks: Assume $Z = 1$ for simplification. Then $\bar{l} \geq D(p||q) + H(U)$. So the best thing we can do is to make $D(p||q)$ as small as possible. By matching $q(u)$ to be similar to p. Suppose that I think that my source is q and I choose a "good" code for q. If the source is really p, then $\bar{l}$ under p is going to have an additional factor that's penalized by $D(p||q)$ which we call the redundancy.

Finally, Shannon code $H(X) \leq \bar{l} \leq H(X) + 1$ for the expected length of a shannon code.

Returning to our earlier example: $l(a) = 1$ and $l(b) = 7$, so $\bar{l} = 1 * 0.99 + 7 * 0.01 = 1.06$ vs. the entropy for this distribution: $H(U) = 0.081$.

Consider instead $U_1, ..., U_N = U^N \in U^N$. Then $P_{U^N}(u^n)$ and the compression rate $\frac{\bar{l}}{N)} \leq \frac{H(U^N)+1}{N} = \frac{N \cdot H(U)+1}{N} = H(U) + \frac{1}{N}$. By coding these up in blocks, the lower bound is smaller (pay a smaller price).

4.2 Huffman Coding

Procedure: almost the same thing as for dyadic sources. e.g.
$p(a) = 0.25$, $p(b) = 0.25$, $p(c) = 0.2$, $p(d) = 0.15$, $p(e) = 0.1$, $p(f) = 0.05$. Goal is to construct a prefix code for this. Recursively merge the two elements with the smallest probabilities until only two things remain. The steps we do in merging this define a binary tree; however, the code is not unique (i.e. if two things have the same probability, you can choose either "branch"). Everytime you make a right in the tree, fill in a '1' if you take a left, fill in a '0' (just keep this consistent).

Theorem 7. *Huffman code is the best possible prefix code. For any other prefix code,*

$$\bar{l}_{CH} \leq \bar{l}_C$$

Proof. Theorem 5.8.1 in Cover and Thomas. □

5 January 25

5.1 Stationary Sources

Goal: moving beyond i.i.d. sources to a more general concept when sequences of random variables are not necessarily i.i.d.

For an optimal code: if $p(u)$ is the probability of the source sequence, then the length of the code for this sequence should approach $\frac{1}{\log p(u)}$.

Now suppose the true distribution follows $p(u)$ but you optimize your codelengths for $q(u)$: $l(u) = \frac{1}{q(u)}$. What happens if you use the wrong distribution to design your code?

$$\begin{aligned}
\mathbb{E}[l(u)] &= \mathbb{E}[\frac{1}{\log q(u)}] \\
&= \mathbb{E}[\log \frac{1}{q(u)} \frac{p(u)}{p(u)}] \\
&= \mathbb{E}[\log \frac{1}{p(u)}] + \mathbb{E}[\log \frac{p(u)}{q(u)}] \\
&= H(P) + KL(p||q)
\end{aligned}$$

The relative entropy is the additional cost that penalizes choosing a non-optimal code by how far the distribution q is from p. Intuitively, if p and q are close, then there's little penalty and the code scheme is near-optimal.

Definition 13 (Stochastic Process)**.** *$\mathcal{X}$ is the alphabet. Then*

$$(X_1, X_2, ..., X_n, ...)$$

is a stochastic process where you can have a different probability distribution for each X_i over the alphabet $\mathcal{X}$ (they are not necessarily i.i.d!).

We define the probability of a sequence of not-necessarily i.i.d. random variables:

$$P((X_1, .., X_n) = (x_1, ..., x_n))$$

Unlike when the values are i.i.d., $P((X_1, X_2, ..., X_n) = (x_1, ..., x_n)) \neq \prod_{i=1}^{n} p(x_i)$.

Definition 14 (Stationary Process). *Main idea: time-invariance. The distribution of X_1 is not that different from X_n.*

$$P((X_1, ..., X_n) = (x_1, .., x_n)) = P((X_{l+1}, X_{l+2}, ..., X_{l+n}) = (x_1, .., x_n))$$

$\forall n$, $\forall l$, $\forall (x_1, ..., x_n) \in \mathcal{X}$. *In other words, shifting the sequence in time does not change the probability distribution. That's why it gets the name "stationary": because the probability distribution is invariant to translations in time. Wherever you start in the sequence, you get the exact same probability distribution.*

As a result

$$\mathbb{E}[X_1] = \mathbb{E}[X_n]$$
$$\mathbb{E}[(X_1 + X_2)] = \mathbb{E}[(X_{10} + X_{11})]$$

since we can set $n = 1$: everything stays the same as you move along the time series.
Examples: *IID fair coin toss:* $\mathcal{X} = \{H, T\}$. $(x_1, ..., x_n) \in \{H, T\}^n$. Then $P((X_1, ..., X_n) = (x_1, ..., x_n)) = P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n) = 2^{-n}$ because each toss is i.i.d. Clearly, this stochastic process is stationary because shifting along the sequence: $P(X_1, ..., X_{n/2}) = P(X_{n/2+1}, ..., X_n)$.

Markov Chain: $P(X_1 = H) = \frac{1}{2}$, $P(X_n = H|X_{n-1} = H) = \frac{3}{4}$, $P(X_n = T|X_{n-1} = T) = \frac{3}{4}$ and X_n is independent of $X_{n-2}, ..., X_1$ given X_{n-1}. This independence is called the "markov property": conditioning on the last realized value makes the current RV independent of all previous RVs.

$$\begin{aligned}P(X_1 = H, X_2 = T, X_3 = H) &= P(X_1 = H)P(X_2 = T|X_1 = H)P(X_3 = H|X_2 = T, X_1 = H)\\ &= \frac{1}{2}\frac{1}{4}P(X_3 = H|X_2 = T)\\ &= \frac{1}{2}\frac{1}{4}\frac{1}{4}\end{aligned}$$

Markov property: only condition on the immediately previous RV realized value.

Arrival times of a bus: $X_1, X_2, ...$ are the arrival times of a bus so that $X_1 - X_2$, $X_3 - X_2$ are iid and positive. Also $\mathbb{E}[X_2 - X_1] = \mu > 0$. The difference in arrival times is greater than 0 and each difference is independent of the last difference. Is this stationary?
Is the distribution of X_1 equal to the distribution of X_2 equal to the distribution of X_3 and so on? Is the distribution $X_1, X_2, ...,$ stationary? Simple check: look at the mean of different RVs throughout time:

$$\begin{aligned}\mathbb{E}[X_2] &= \mathbb{E}[X_1 + X_2 - X_1]\\ &= \mathbb{E}[X_1] + \mathbb{E}[X_2 - X_1]\end{aligned}$$

so this stochastic process is not stationary as the mean is increasing $\mathbb{E}[X_2 - X_1] > 0$.

However, the stochastic process $X_2 - X_1, X_3 - X_2, ...$ is stationary because each element is iid. So we can transform a non-stationary stochastic process into a stationary one.

Delta Coding: $(X_1, ..., X_n, ...)$ not stationary distribution → (Δ-coding invertible transformation) $(X_1, (X_2 - X_1), (X_3 - X_2), ...)$ becomes stationary.

Many compression algorithms require stationary processes as otherwise they do not compress very well.

Text: If you take an english book and start looking at word-by-word (each X_i is a word). Is this stationary or non-stationary? If you read the first page of a book and the last page, then you're able to distinguish between them. However, to a first-approximation, it is stationary: sequence of words on the nth page look roughly similar to the $n + 1$th page. If you take the expectation of the k^{th} word, this should roughly equal the expectation of the $(k + 1)^{th}$ word of the book. However, in reality, the word "the" is more likely to occur in the first word rather than the k^{th} word. However, this is good enough to a first approximation.

5.2 Entropy Rate:

For the entropy of an iid stochastic process $H(X_1, ..., X_n)$, we could simply compute the entropy of $H(X_1)$ and then multiply by n to compute the entropy of this sequence since everything is independent.

Definition 15 (Entropy Rate of a Stationary Process)**.** *Simply, entropy rate is the average entropy per random variable in your stochastic process, or the "average entropy in the process". Therefore, for a non-iid stochastic process*

$$\lim_{n\to\infty} \frac{H(X_1, ..., X_n)}{n}$$

we need to take the limit as $n \to \infty$ as we can have very far dependencies between X_i and X_{i+k} so that this makes sense for every stationary process.

Another way to think about the entropy rate:

$$\lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_1)$$

as how many bits do I need to store the n^{th} symbol given I have already comperssed all the symbols before it. How much entropy does the n^{th} symbol introduce to this process.

This is the entropy of the $X_n th$ RV given the ones before it for as $n \to \infty$

Theorem 8. *For a stationary process, $\lim_{n\to\infty} \frac{H(X_1,...,X_n)}{n}$ and $\lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_1)$ both exist and moreover*

$$\lim_{n\to\infty} \frac{H(X_1, ..., X_n)}{n} = \lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_1)$$

with $H(\mathcal{X}) :=$ the entropy rate.

Returning to our coin-toss example, we can easily show via the markov property.

$$H(X_n|X_{n-1}, ..., X_1) = H(X_n|X_{n-1})$$

But can we transform this sequence $(X_1, ..., X_n)$ so that the sequence is now iid? Define new RV $Y_i = \begin{cases} 1 if X_i = X_{i-1} \\ 0 \text{ otherwise.} \end{cases}$. Then $H, T, H, H \Rightarrow H, 0, 0, 1$, and then we can show the Y_i's are independent.

5.3 AEP for Stationary Processes

$$\frac{-1}{n} \log p(x_1, ..., x_n) \to H(\mathcal{X})$$

with probability 1. Recall $H\mathcal{X}))$ is the entropy rate. Moreover, we previously showed that the entropy rate is the lower bound on the expected length for any uniquely decodable code. This is maintained for stationary processes: entropy rate remains the lower bound.

5.4 Universal Compressors: LZ77

Universal compression: you have a compression scheme that will work well for *any* distribution (you don't know it ahead of time).

In real life Huffman coding is not good enough because our sources are not iid.

Consider a compressor C with length function $l(X^n)$ is called a universal compressor if

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}[l(X^n)] = H(\mathcal{X})$$

for every stationary source distribution. The idea is that as n becomes large, the expected length becomes optimal.

LZ77: Ziv & Lempel 1977 is a universal compressor.

Core idea: if you have any sequence that occurs with very high probability, then you see it very often. Instead of storing these high frequency sequences, you store a pointer to the last time you saw this sequence (rather than storing the entire codeword/string for this high-frequency sequence). This gives us the optimal in the limit.

This is called dictionary coding. When you encounter a string, you search the dictionary for the last time you saw it, and store a pointer to this last time. Formally:

Input: Sequence $x_1, x_2, ..., x_n, ...$

Suppose we're at x_i. We find the largest k such that for some $j < i$, $(x_j, ..., x_{j+k-1}) = (x_i, ..., x_{i+k-1})$ [finding the largest exact match]. If you find a match, then store $(i-j, k, x_{i+k})$ (how far ago was the match, the length of the match, and the first unmatched symbol). If no match, store $(0, 0, x_i)$.

Example: ABBABBABBBAA. We store:

1. (0,0,A)
2. (0, 0, B)
3. (1, 1, A)
4. (3, 5, B)
5. (4,1,A)

Then we can decode this sequence:

1. (0,0,A) → A
2. (0,0,B) → B
3. (1,1,A) → BA
4. (3,5,B) → BBABBB
5. (4,1,A) → AA

Theorem 9. *LZ77 is universal.*

gzip uses LZ77 and then Huffman encoding to encode the 3-tuples. 7-zip uses LZ77 followed by Arithmetic coding.

5.5 General Tips on Compression

The limit is equal to infinity, but in reality, we always have a finite n. Therefore, LZ77—while universal—applied on real data may not be optimal.

Making your data more iid often helps a lot in terms of compression.

In reality, you have a complex data source, you convert it into multiple data streams (each of which are stationary or iid), and then you pass each stream to gzip. This speeds up compression + uses window size for how far LZ77 looks back.

You can also approximate the entropy of a source. If gzip is giving you 1 GB and the approximate entropy is 0.99 GB, then it's likely a waste of time to further work on optimizing your source.

6 January 30

For a sequence of conditional probabilities $\{q(x_t|x^{t-1})\}_{t=1}^n$ this is equivalent to finding compression for this sequence. $\sum_{t=1}^n \log \frac{1}{q(x_t|x^{t-1})} = \log \frac{1}{q(x^n)}$ [this is called log loss or perplexity].

Kraft inequality: for any compressor of sequences of length n, $\sum_{x^n} 2^{-l(x^n)} \leq 1$. Moreover, ≈ 1 for good compressors. Moreover, it induces a probability mass function where $q(x^n) = 2^{-l(x^n)}$.

Cool interplay between compression problems and learning problems. We try to compress distributions by using a deep neural network to model $q(x_t|x^{t-1})$ and then feeding these values into a compressor (i.e. arithmetic compressor).

Open question: can we use compression to estimate $q(x_t|x^{t-1})$ that's used to augment deep learning or replace these conditional probabilities.

6.1 Noisy Channel

X → Noisy Channel → Y.
Noisy Channel is simply $P(Y|X)$. Given some channel input, there's some distribution on the channel outputs.

Example 2 (Binary Symmetric Channel)**.** *Also denoted as BSC and BSC(δ). $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$ are both binary and $P_{Y|X}(y|x) = \begin{cases} \delta \text{ if } x \neq y \\ 1-\delta \text{ if } x = y \end{cases}$ Another way to represent this is $Y = X \bigoplus N$ where X, N are independent and $N \ \sim Ber(\delta)$ is the noise added to X.*

Example 3 (Binary Erasure Channel)**.** *Also denoted as $BEC(\delta)$. $\mathcal{X} = \{0,1\}$ and $Y = \{0,1,e\}$ where we denote a symbol e for "erased" where the information is lost. Each $1,0$ has probability δ as getting erased.*

Probability of being erased independent of whatever the input is.

$$\begin{aligned} H(X|Y) &= H(X|Y=e)P(Y=e) + H(X|Y=0)P(Y=0) + H(X|Y=1)P(Y=1) \\ &= H(X) * P(Y=e) \\ &= H(X)\delta \end{aligned}$$

Where we use $H(X|Y=e) = H(X)$, $H(X|Y=1) = H(X|Y=0) = 0$ since we know exactly what the input was given the output. So this conditional entropy is simply H(X) δ.

6.2 Informational Capacity

Look at the mutual information between the input and output of the channel: $I(X;Y)$. Mutual information is a natural measure of how one Random Variable is important to another. We will maximize it over all the possible distributions of the input (since the channel is specified as a conditional $P(Y|X)$):

$$C^{(I)} := \max_{P_X} I(X;Y)$$

where $C^{(I)}$ is defined as the informational capacity of a channel.

For our eraser example with $H(X|Y) = H(X)\delta$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= (1-\delta)H(X) \end{aligned}$$

⇒ the uniform distribution will maximize $I(X;Y)$.

$$\begin{aligned} C^{(I)} &= \max_{X} I(X;Y) \\ &= (1-\delta)H(X) \end{aligned}$$

Where the maximum is achieved by the input distribution $X \sim Ber(0.5)$.

For BSC(δ), [recall $\bigoplus$ stands for addition mod 2]

$$\begin{aligned} H(Y|X) &= H(X \bigoplus N|X) \\ &= H(N) \\ &= H_2(\delta) \end{aligned}$$

So to maximize the mutual information between X and Y:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H_2(\delta) \end{aligned}$$

we have:

$$\begin{aligned} C^{(I)} &= \max_{X} I(X;Y) \\ &= \max_{X} H(Y) - H_2(\delta) \\ &= 1 - H_2(\delta) \end{aligned}$$

which is achieved by $X \sim Ber(0.5)$. The channel capacity asks which distribution on the channel input (i.e. what distribution on X) will maximize the entropy on Y? The answer is a uniform distribution. If $X \sim Ber(0.5)$, then $Y \sim Ber(0.5)$. And Bernoulli 0.5 is the maximum entropy you'd get for any random variable.

6.3 Communication over a Noisy Channel

m bits → Encoder [or transmitter] → $X_1, X_2, ..., X_n$ → Noisy Channel → $Y_1, Y_2, ..., Y_n$ → Decoder [or receiver] that tries to recreate the m bits given to the encoder.

We assume a memory-less channel:

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$$

statistical relationship between one channel input and one channel output [for a memoryless channel]. This presents a notion if independence: the noise usage across this channel is independent for each random variable. The flip events are iid.

Figures merit. For now assume the m-bits are iid$\sim Ber(0.5)$. In other words, there's a message J uniformly distributed on a set of size 2^m: $\{1, 2, ..., 2^m\}$.

1. Reliability of the system: $P_e = P(\text{m bits} \neq \text{m bits})$ as the probability of an error: $P(J \neq \hat{J})$. The probability that we input m bits and get out a number of bits other than m.

2. Rate of communication: $\frac{m}{n}$ where m is the number of bits and n is the number of times you use the channel.

Det R is "an achievable rate for reliable communication" or "achievable" (for short) if you can communicate R bits per channel use reliably (i.e. with error that can be arbitrarily small). More formally,
$\forall \epsilon > 0$, $\exists$ (m,n, encoder, decoder) (number of bits, number of channel usages, encoder, decoder) with $P_e \leq \epsilon$ and $\frac{m}{n} \geq R$.

The channel capacity $C :=$ sup of achievable rates (achievable rates in terms of reliable communication). This is the maximal rate of reliable communication (per channel use).

Noisy channel coding theorem (Shannon 1948): $C = C^{(I)} = \max_{P_X} I(X;Y)$. Converse part $C \leq C^{(I)}$ and we can also prove the direct part (constructive part): $C \leq C^{(I)}$.

7 February 1

Shannon 1948: For a memory-less channel characterized by $P_{Y|X}$: we have a capacity C (maximum number of bits for channel use that you can communicate reliably with low probability of error).
$C = \max_X I(X;Y)$

E.g. $BSC(\delta)$: binary input, binary output, and flips the bit through the channel with probability δ. $BSC(\delta) : C = 1 - h_2(\delta)$ where h_2 is the binary entropy of probability δ.

Consider the set of possible 2^n channel outputs. You send in $X^n(j)$ and you get out a sequence that lies in 2^{nH} by the AEP: "noise ball" around your message $X^n(j)$ that is of size $2^{nh_2(\delta)}$. If you send a certain sequence that corresponds to a message j, given that, you know wiht all likelihood, what's going to come out of the channel output will be somewhere in the "typical noise ball" of radius δ from AEP. This set will have size $2^{nh_2(\delta)}$. This is the typical noisy outputs around the input you send into the channel.

The number of messages that we can hope to communicate reliably, we need the "typical noise balls" to non-intersect for different input messages. The number of possible messages cannot be greater than the size of all possible channel inputs divided by the size of a typical noise ball (otherwise, will get intersection between noise balls from input messages):
number of messages $\leq \frac{2^n}{2^{nh_2(\delta}} = 2^{n(1-h_2(\delta)}$
So the number of bits (log number of messages) $\leq n(1 - h_2(\delta))$ [communicating m bits $\iff$ communicating 2^m possible messages. And communicating $\log m$ bits $\iff$ communicating m messages].
And the channel use: or the number of bits divided by n $\leq (1 - h_2(\delta))$ and this is indeed the capacity of $BSC(\delta)$.

The intuitive way to think about Capacity for BSC(δ) is that you can't do any better than $(1 - h_2(\delta))$ from a counting/volume perspective. One thing we're using is an AEP/law of large numbers argument that the output sequence will–with high probability–be concentrated in a small subset of all binary sequences.

When you work with high-dimensions, you can be very effective to get no overlap in noise balls, but the packing is very effective. You can effectively pack $2^{n(1-h_2(\delta))}$ messages without noise-ball overlap.

Same intuition for a general channel: $I(X;Y) = H(Y) - H(Y|X)$.First term: exponential growth of channel inputs - (second term) Corresponds to the exponential behavior of channel outputs given channel inputs. What to make the set of channel outputs as big as it can be.

7.1 Measures of Information For Continuous Random Variables

Definition 16 (Relative Entropy)**.** *The relative entropy between two probability density functions (PDFs) f and g:*

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)}$$

Definition 17 (Mutual Information)**.** *The mutual information between continuous random variables X and Y that have a joint PDF $f_{X,Y}$ is*

$$I(X,Y) = D(f_{X,Y}||f_X \times f_Y)$$

where $D(f_{X,Y}||g_{X,Y}) = \int\int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{g_{X,Y}(x,y)} dxdy$

Definition 18 (Differential Entropy)**.** *The Differential Entropy of a continuous random variable X with PDF f_X is*

$$\begin{aligned} h(x) &:= \mathbb{E}[-\log f_x(X)] \\ &= -\int f_X(x) \log f_X(x) dx \end{aligned}$$

We call it differential rather than simply "entropy" because

Definition 19 (Joint and Conditional Differential Entropy)**.** *If X, Y have a joint density $f_{X,Y}$, the conditional differential entropy is*

$$h(X|Y) := \mathbb{E}[-\log f_{X|Y}(X|Y)]$$

and the joint differential entropy is

$$h(X,Y) := \mathbb{E}[-\log f_{X,Y}(X,Y)]$$

Property 7.1.1.

$$h(X|Y) \leq h(X)$$

with equality iff X and Y are independent.

Property 7.1.2.

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y) \end{aligned}$$

Property 7.1.3. *For a constant a*

$$h(X+a) = h(X)$$

Differential entropy is invariant to translation.

Property 7.1.4. *For a constant $a \neq 0$*

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

If a is sufficiently small $\log |a|$ *can be highly negative, and differential entropy itself can be negative.*

Example 4 (Gaussian Distribution)**.** *Let* $X \sim N(\mu, \sigma^2) \iff f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}x^2}$. *Then because* $h(X + a)$ *is invariant to shifts, we can assume* $\mu = 0$.

$$\begin{aligned}
h(X) &= \mathbb{E}[-\log f_X(X)] \\
&= \frac{1}{\ln 2}\mathbb{E}[-\ln f_X(X)] \\
&= \frac{1}{\ln 2}\mathbb{E}[\frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}x^2] \\
&= \frac{1}{2\ln 2}[\ln(2\pi\sigma^2) + \frac{\sigma^2}{\sigma^2}] \\
&= \frac{1}{2\ln 2}[\ln(2\pi\sigma^2) + \ln e] \\
&= \frac{1}{2\ln 2}[\ln(2\pi\sigma^2 e)] \\
&= \frac{1}{2}\log(2\pi\sigma^2 e)
\end{aligned}$$

8 February 6

8.1 Channel Capacity with Constraints

$J \in \{1, 2, ..., M\}$ all equally probable. We encode this into $X_1, ..., X_n$, send it through a memoryless channel $P_{Y|X}$ to get out $Y_1, ..., Y_n$ that the receiver then decodes to $\hat{J}$.

Recall the rate is $\frac{\log m}{n}$ and the probability of error $P_e(\hat{J} \neq J)$ is the probability that an error has been made.We want to reconstruct the input exactly.

We define the channel capacity as the supremum over all achievable rates for reliable communication [maximum rate such that P_e is arbitrarily small].

Shannon 1948: $C = \max_{P_X} I(X;Y)$: the channel capacity is equal to the mutual information between X and Y when this quantity is maximized over P_X.

Note: Sometimes we also have a cost constraint in the form

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i) \leq \alpha$$

The average probability does not exceed α or another constraint:

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \rho(X_i)] \leq \alpha$$

$C(\alpha)$ is the channel capacity under cost constraint α. $C(\alpha)$ is defined like C when only schemes satisfying either of the above cost constraints are allowed.

Shannon's theorem carries over:

$$C(\alpha) = \max_{X|\mathbb{E}[\rho(X)] \leq \alpha} I(X;Y)$$

8.2 Differential Entropy

If X is a continuous RV with f_X, then its **differential entropy** is

$$h(X) = \mathbb{E}[\log \frac{1}{f_X(X)}]$$

If $G \sim N(0, \sigma^2)$, then $h(G) = \frac{1}{2}\log(2\pi e\sigma^2)$. If σ^2 is large enough, this will be negative. Therefore, differential entropy does not have any of the same significance that entropy does. However, we can use it for computing mutual information. Gaussian is actually somewhat special – among all RVs with the same variance, a Gaussian has maximum differential entropy.

Theorem 10. *Suppose X is a random variable with $\mathbb{E}[X^2] \leq \sigma^2$ [second moment is less than σ^2]. Then*

$$h(X) \leq h(G)$$

with equality $\iff$ $X \sim N(0, \sigma^2)$

Proof.

$$\begin{aligned}
0 &\leq D(f_X||f_G) \\
&= \int f_X(x)\log\frac{f_X(x)}{f_G(x)}dx \\
&= \mathbb{E}[\log\frac{f_X(X)}{f_G(X)}] \\
&= -h(X) + \mathbb{E}[\log\frac{1}{f_G(X)}] \\
&= -h(X) + \mathbb{E}[-\log f_G(X)] \\
&= -h(X) + \mathbb{E}[-\log\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}X^2}] \\
&= -h(X) + \mathbb{E}[-\log\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{\ln 2}\frac{X^2}{2\sigma^2}] \\
&= -h(X) + -\log\frac{1}{\sqrt{2\pi\sigma^2}} + \mathbb{E}[\frac{1}{\ln 2}\frac{X^2}{2\sigma^2}] \\
&\leq -h(X) + -\log\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{\ln 2}\frac{\sigma^2}{2\sigma^2} \\
&\text{used } \mathbb{E}[X^2] \leq \sigma^2 \\
&= -h(X) + \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\frac{\ln e}{\ln 2} \\
&= -h(X) + \frac{1}{2}\log(2\pi\sigma^2 e) \\
&= -h(X) + h(G)
\end{aligned}$$

So we've found that the entropy of X is always less than or equal to Y. This inequality becomes equality $\iff$ X is distributed as Y since the two inequalities turn into equalities when this occurs. $D(f_X||f_G) = 0$ when $f_X = f_G$ and moreover, $\mathbb{E}[X^2] = \sigma^2$ when $X \sim N(0, \sigma^2)$. □

Example 5. *AWGN [additive white Gaussian noise] channel —really additive white memoryless channel: $Y_i = X_i + N_i$ where N_i is iid with $N_i \sim N(0, \sigma^2)$.*

Say we have a power constraint:

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2] \leq p$$

Shannon: $C(P) = \max_{\mathbb{E}[X^2]\leq P} I(X;Y)$ with $Y = X + N$, $N \sim (0, \sigma^2)$ independent of X.

For any random variable X with $\mathbb{E}[X^2] \leq p$

$$\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) \\
&= h(Y) - h(Y - X|X) \\
&\text{X is deterministic given } X\text{, so you can look at Y-constant − invariant given a shift.} \\
&= h(Y) - h(N|X) \\
&= h(Y) - h(N)
\end{aligned}$$

$\mathbb{E}[Y^2] = \mathbb{E}[(X+N)^2] = \mathbb{E}[X^2] + \mathbb{E}[N^2] + 2\mathbb{E}[XN]$ and $2\mathbb{E}[XN]$ since N has mean 0, so $\mathbb{E}[X^2] + \mathbb{E}[N^2] \le P + \sigma^2$ so Y will be upper-bounded by a gaussian with variance $P + \sigma^2$:

$$\begin{aligned} h(Y) - h(N) &\le \frac{1}{2}\log(2\pi e(P+\sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) \\ &= \frac{1}{2}\log(1 + \frac{p}{\sigma^2}) \end{aligned}$$

So for any random variable X with second moment less than p, we know the mutual information will be upper-bounded by $\frac{1}{2}\log(1+\frac{p}{\sigma^2})$. Question: is there any distribution in the feasible set (i.e. second moment constraint) where the one inequality holds with equality? Yes: if $X \sim N(0,p)$, then $Y \sim N(0, p+\sigma^2)$ and $I(X;Y) = \frac{1}{2}\log(1+\frac{p}{\sigma^2})$ because $\mathbb{E}[Y^2] = p + \sigma^2$.

Therefore, $C(P) = \frac{1}{2}\log(1 - SNR)$ where SNR $= \frac{p}{\sigma^2}$ = the number of bits per channel you can reliable communicate over an additive white Gaussian noise channel.

$$\begin{aligned} \mathbb{E}[Y^2] &\le \mathbb{E}[\sum_{i=1}^{n} X_i^2 + \sigma^2] \\ &= \mathbb{E}[\sum_{i=1}^{n} X_i^2] + n\sigma^2 \\ &\le n(p+\sigma^2) \end{aligned}$$

using that $\mathbb{E}[\sum_{i=1}^{n} X_i^2] \le np$ [power constraint]

Using noise-ball perspective: how many small balls can you pack into the big ball. Get into trouble when two small balls intersect: do not know which message to decode this output to. So the number of messages $\le$ Volume of the ball in $\mathbb{R}^n$ of radius $\sqrt{n(p+\sigma^2)}$ divided by the total volume with radius $\sqrt{n(p+\sigma^2)}$ [due to the constraint]. So we have:
number of messages $= \frac{k_n(\sqrt{n(p+\sigma^2)})^n}{k_n(\sqrt{n\sigma^2})^2} = (1+\frac{p}{\sigma^2})^{\frac{n}{2}}$ for reliable communication. Cannot hope to pack more noise balls in the circular volume of $X^n(j)$ specified by the constraint.

Therefore, $\frac{1}{n}$ log of the number of messages [rate of communication] $\le \frac{1}{n}\log(1+\frac{p}{\sigma^2})^{\frac{n}{2}} = \frac{1}{2}\log(1+\frac{p}{\sigma^2})$.

9 February 8

9.1 Fano's Inequality

X is a Random Variable with alphabet $\mathcal{X} = \{1, ..., M\}$. Fix $i \in \mathcal{X}$ and consider the entropy of X (entropy of X is equal to entropy of X + any deterministic function of X). Let $1_{\{X=i\}}$ be the indicator for whether $X = i$ (or not)

$$\begin{aligned} H(X) &= H(X, 1_{\{X\neq i\}}) \\ &= H(1_{\{X\neq i\}}) + H(X|1_{\{X\neq i\}}) \\ &= h_2(P(X\neq i)) + H(X|X\neq i)P(X\neq i) + H(X|X=i)P(X=i) \end{aligned}$$

but note $H(X|X=i) = 0$ because this is deterministic. And $H(X|X\neq i) \le \log(M-1)$ because a uniform distribution has the highest entropy.

$$\begin{aligned} h_2(P(X\neq i)) + H(X|X\neq i)P(X\neq i) + H(X|X=i)P(X=i) &= h_2(P(X\neq i)) + H(X|X\neq i)P(X\neq i) \\ &\le h_2(P(X\neq i)) + P(X\neq i)\log(M-1) \end{aligned}$$

Now let Y be a Random Variable (which may or may not be correlated with X), and let $\hat{X} = g(Y)$. Interpret $\hat{X}$ as an attempt to guess the value of X as a function of Y. Let $P_e = P(X \neq \hat{X})$ be the probability that our guess was incorrect.

$$H(X|Y) = \sum_{y} H(X|Y=y)P(Y=y)$$

We can now upper-bound each $H(X|Y=y)$ by $h_2(P(X \neq i)) + P(X \neq i)\log(M-1)$ when the role of i is being played by $i = g(y)$. Specifically, $X|Y=y$ is a random variable so we can bound each one by the earlier inequality:

$$\sum_y H(X|Y=y)P(Y=y) \leq \sum_y [h_2(P(X \neq g(Y))|Y=y) + P(X \neq g(Y)|Y=y)\log(M-1)]P(Y=y)$$

Recall h_2 is a concave function of the input, so it's upperbounded by $h_2(\sum_y P(X \neq g(Y)|Y=y))P(Y=y)$ [Consequence of Jensen's]. So we have

$$\begin{aligned}\sum_y [h_2(P(X \neq g(Y))|Y=y) + P(X \neq g(Y)|Y=y)\log(M-1)]P(Y=y) &\leq h_2(\sum_y P(X \neq g(Y)|Y=y))P(Y=y) + \sum_y P(X \neq g(Y)|Y \\ &= h_2(P(X \neq g(Y))) + P(X \neq g(Y))\log(M-1) \\ H(X|Y) &\leq h_2(P_e) + P_e \log(M-1)\end{aligned}$$

where $h_2(\sum_y P(X \neq g(Y)|Y=y))P(Y=y) = h_2(P(X \neq g(Y)))$ by the law of total probability and $\sum_y P(X \neq g(Y)|Y=y)P(Y=y)\log(M-1) = P(X \neq g(Y))\log(M-1)$ again by the total probability.

$H(X|Y) \leq h_2(P_e) + P_e \log(M-1)$ is Fano's inequality.

9.2 Converse Part of Shannon's Theorem

Communications Setting: $J \in \{1, ..., M\}$ and $P(J=i) = \frac{1}{M}$ we have M different messages each of which are equiprobable of begin communicated.

$J \in \{1, ..., M\} \Rightarrow$ Encoder $X_1, ..., X_N \Rightarrow$ Memoryless Channel $P_{Y|X} \Rightarrow Y_1, ..., Y_n \Rightarrow$ decoder $\Rightarrow \hat{J}$. We define $P_e = P(J \neq \hat{J})$ and rate of communication is $\log \frac{M}{n}$ which is the number of bits communicated divided by the number of channel usages.

Taking for granted that you can communicated reliably over noisy channels with positive rates: $C = C^{(I)} := \max_X I(X;Y)$.

Theorem 11 (Converse Part of Shannon's Theorem)**.**

$$C^{(I)} \leq \max_X I(X;Y)$$

Proof. Fix $R > C^{(I)}$ and any scheme with rate $\frac{1}{n}\log M \geq R$. Recall $J \sim U$, so $H(J) = \log M$. Consider:

$$\begin{aligned}\log M - H(J|Y^n) &= H(J) \\ &= H(J) - H(J|Y^n) \\ &= I(J;Y^n) \\ &= H(Y^n) - H(Y^n|J) \\ &= \sum_{i=1}^n H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, J) \\ &\leq \sum_{i=1}^n H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, J, X_i) \\ &= \sum_{i=1}^n H(Y_i|Y^{i-1}) - H(Y_i|X_i)\end{aligned}$$

because the channel is memoryless, $H(Y_i|Y^{i-1}, J, X_i) = H(Y_i|X_i)$ – no longer matters what the last channel output was. We have the Markov Chain (Markov-Triplet: X-Y-Z; X and Z are independent given Y): $Y_i - X_i - (J, Y^{i-1})$ because the channel is memoryless.

$$\begin{aligned}\sum_{i=1}^n H(Y_i|Y^{i-1}) - H(Y_i|X_i) &\leq \sum_{i=1}^n H(Y_i) - H(Y_i|X_i) \\ &= \sum_{i=1}^n I(X_i;Y_i) \\ &\leq \sum_{i=1}^n \max_X I(X_i, Y_i) \\ &\leq nC^{(I)}\end{aligned}$$

So rearranging, we have $H(J|Y^n) \geq \log M - nC^{(I)}$. So you can't hope to reduce the initial uncertainty on the initial message by more than $C^{(I)}$ for each additional channel use. We want to relate this to the probability of error: does sending the same message more times reduce P_e?

A weaker version of Fano's Inequality: $H(X|Y) \leq 1 + P_e \log M$ [binary entropy upperbounded by 1]. $\Rightarrow P_e \geq \frac{H(X|Y)-1}{\log M}$
Returning to $H(J|Y^n) \geq \log M - nC^{(I)}$:

$$\begin{aligned} P_e &\geq \frac{H(J|Y^n)-1}{\log M} \\ &\geq \frac{\log M - nC^{(I)} - 1}{\log M} \\ &= 1 - \frac{n}{\log M}C^{(I)} - \frac{1}{\log M} \\ &\geq 1 - \frac{1}{R}C^{(I)} - \frac{1}{\log M} \\ &\geq 1 - \frac{1}{R}C^{(I)} - \frac{1}{\log nR} \end{aligned}$$

using $\frac{n}{\log M} \leq \frac{1}{R}$ from earlier in the problem statement. As $n \to \infty$ [i.e. we use the channel many times], we have

$$\lim_{n\to\infty} 1 - \frac{1}{R}C^{(I)} - \frac{1}{\log nR} = 1 - \frac{C^{(I)}}{R} \qquad > 0$$

because $R > C^{(I)}$. So the Probability of error cannot go to 0. Any sequence of schemes with rates $\geq R$ (rate of communication larger than channel capacity) cannot have associated probability of error P_e that are vanishing (i.e. $P_e \to 0$ as $n \to 0$) if $R > C^{(I)}$. □

Note: The proof we just completed carries over verbatim to $X_i = X_i(J, Y^{i-1})$ [i.e. this setting is feedback communication – what you feed into the channel in the i^{th} time step can depend on everything that came out of the channel so far].
⇒ Feedback does not increase channel capacity for a memory-less channel!

Recall the binary erasure channel (BEC): input is binary $\in \{0, 1\}$ and output may be erased with some probability: $\{0, 1, e\}$ erased is $C^{(I)} = 1 - \delta$. With feedback, we can send a bit again if it got erased!

Consider scheme that repeats every information bit until success (i.e. it goes through unerased). How many channel uses on-average will I need to send a bit so that it goes through without getting erased? $\frac{1}{1-\delta}$ because it's follows a Geometric distribution. Expected value of a Geometric Distribution with probability $1 - \delta$ [probability of success]. Therefore $\frac{1}{1-\delta}$ channel uses per each information bit. Alternatively, in terms of communication rate—i.e. how many information bits per channel bits—we're getting $1 - \delta$ bits per channel use.

What is my probability of error for this system? $P_e = 0$ because there's no chance of making an error: bits go through error free. The moral of the story is that feedback does not increase capacity, but it can very significant improve reliability and the simplicity of schemes that do achieve capacity. It remains to show that even without feedback, you can achieve channel capacity.

10 February 13

Communication setting [Noisy Channel]:
$J \in \{1, 2, ..., M\} \to$ Encoder $\to X_1, ..., X_n \to$ Noisy Channel $P_{Y|X} \to$ Decoder $\hat{J}$.

We look at two different performance criteria:

1. The probability of an error: $P_e = P(\hat{J} \neq J)$
2. The rate of the scheme: $\frac{\log M}{n}$ or the number of bits per channel use.

Encoder is really equivalent to a choice of mapping from Message to Sequence of Channel inputs that I will send or transmit. Encoder $\Longleftrightarrow$ $C_n = \{X^n(1), X^n(2), ..., X^n(M)\}$ set of M possible sequences that I will send to transmit the message. We denote C_n as the choice of the codebook.

Today: We'll prove the "direct" part: $\forall P_X$ and any rate $R \leq I(X;Y)$, if $R \leq I(X;Y)$, then necessarily R is achievable. Achievable: there is a sequence of schemes with at least rate R with probability of error that is vanishing.

Joint AEP: Today we will extend the AEP to a pair of random variables: $(X, Y) \sim P_{X,Y}$ with $\mathcal{X}$, $\mathcal{Y}$ finite alphabet. Assume pairs (X_i, Y_i) are i.i.d. $\sim (X, Y)$ as usual: $P(X^n) = \prod_{i=1}^n P_X(X_i)$ and $P(Y^n) = \prod_{i=1}^n P_Y(Y_i)$.

So $P(X^n, Y^n) = \prod_{i=1}^n P_{X,Y}(x_i, y_i)$ because we have a joint iid pmf.

Define: $A_\epsilon^{(n)}(X, Y)$ as the typical set associated with a pair of random variables (X, Y).

$$A_\epsilon^{(n)} = \{(x^n, y^n) \text{ s.t. } |\frac{-1}{n}\log P(X^n) - H(X)| \leq \epsilon, |\frac{-1}{n}\log P(Y^n) - H(Y)| \leq \epsilon, |\frac{-1}{n}\log P(X^n, Y^n) - H(X, Y)| \leq \epsilon\}$$

The X-sequence is typical, the Y-sequence is typical, and the sequence of X, Y-pairs is typical.

Theorem 12. *$\forall \epsilon > 0$, $P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$*

Moreover, $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}(X, Y)| \leq 2^{n(H(X,Y)+\epsilon)}$

Proof. By the AEP that we already know, the sequence X^n will satisfy the condition. □

Suppose $\tilde{X}^n = X^n$ (distributed i.i.d. according to X^n) and $\tilde{Y}^n = Y^n$ are independent:

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X, Y)) \to 0$$

as $n \to \infty$

1. $\tilde{X}^n$ distributed i.i.d. uniformly on $A_\epsilon^{(\epsilon)}(X)$.
2. Similarly, $\tilde{Y}^n$ is essentially uniformly distributed on the typical set associated with Y: $A_\epsilon^{(n)}(Y)$
3. From (1) and (2) the pair $(\tilde{X}^n, \tilde{Y}^n)$ is uniform distributed on the set $A_\epsilon^{(n)}(X) \times A_\epsilon^n(Y)$ [Cartesian product of sets where the first component is in $A_\epsilon^{(n)}(X)$ and second component is in $A_\epsilon^{(n)}(Y)$].
4. $P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X, Y)) \approx \frac{|A_\epsilon^{(n)}(X,Y)|}{|A_\epsilon^{(n)}(X)||A_\epsilon^{(n)}(Y)|} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-n[H(X)+H(Y)-H(X,Y)]} = 2^{-nI(X,Y)}$ because the joint AEP is a smaller subset of the cartesian product of the individual X and Y AEPs: we can show the size of the joint set smaller than the product set.

It's exponentially unlikely that these two random variables are jointly typical.

Theorem 13. *For any $\epsilon > 0$ and n sufficiently large, the probability of two sequences that are independent (marginally independent)*

$$(1-\epsilon)2^{-n(I(X,Y)+3\epsilon)} \leq P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X, Y)) \leq 2^{-n(I(X,Y)-3\epsilon)}$$

For a pair $(X^n(J), Y^n)$ where Y^n is the output after sending X^n through the memoryless noisy channel, then with high probability, $(X^n(J), Y^n) \in A_\epsilon^{(n)}(X, Y)$. However, for any $k \neq J$, $(X^n(k), Y^n) \in A_\epsilon^{(n)}(X, Y)$ is exponentially small according to the mutual information between X, Y. For any message that is not sent, it's codeword is independent from the message that is sent, and therefore Y^n is independent from $X^n(k)$. So these are two i.i.d. independent sequences so the probability these two sequences are jointly typical is $\approx 2^{-nI(X;Y)}$ for any other possible message $k \neq J$.

Joint-Typicality Decoder: take $\hat{J}$ such that $(X^n(\hat{J}), Y^n) \in A_\epsilon^{(n)}(X, Y)$ (i.e. is jointly typical with the channel output). This decoder will get you $J = \hat{J}$ with high probability provided the size of the codebook is $\leq 2^{nR}$ where $R \leq I(X; Y)$. Exponentially low probability you confuse $X^n(k) \to Y^n$ with the true message: $X^n(J) \to Y^n$.

Proof. Direct part of the channel coding theorem: Fix P_X and $R < I(X; Y)$. Need to show that R is an achievable rate for reliable communication.

Let $M = \lceil 2^{nR} \rceil$ be the size of the codebook. And generate codebook C_n by letting $X^n(1), ..., X^n(M)$ be iid $\sim P_X$.

Consider the decoder: $\hat{J}(Y^n)$ is a function of the output sequence Y^n

$$\hat{J}(Y^n) = \begin{cases} j \text{ if } (X^n(j), Y^n) \in A_\epsilon^{(n)} \text{ and } (X^n(k), Y^n) \in A_\epsilon^{(n)} \ \forall k \neq j \\ \text{error otherwise} \end{cases}$$

We now examine $P_e(C_n)$ (the probability of error associated with this decoder). Notie C_n is a random object (randomly generated codebook), so $P_e(C_n)$ is a random variable. So we can average over all possible C_n to find the expectation of the probability of error given a randomly generated codebook:

$$\begin{aligned}\mathbb{E}[P_e(C_n)] &\le P(J \neq \hat{J})\\ &= \sum_{j=1}^{M} P(J \neq \hat{J}|J = j)P(J = j)\end{aligned}$$

Our scheme is doing something (not the MLE on the conditional likelihood which is the best we can do). Moreover, we're going to get an error if either the received output sequence Y^n is not typical with the codeword associated with the input message $J = j$: i.e., $(X^n(j), Y^n) \notin A_n^{(e)}(X;Y)$ given $J = j$ plus $P((X^n(k), Y^n) \in A_n^{(e)}(X;Y))$ for some $k \neq j$ given message $J = j$ was sent.

$$\begin{aligned}P(J \neq \hat{J}|J = j) &\le P((X^n(j), Y^n) \notin A_n^{(e)}(X;Y)|J = j) + P((X^n(k), Y^n) \in A_n^{(e)}(X;Y)|J = j)\\ &= P((X^n, Y^n) \notin A_n^{(e)}(X;Y)) + P((X^n(k), Y^n) \in A_n^{(e)}(X;Y) \text{ for some k} \neq j \ |J = j)\\ &= P((X^n, Y^n) \notin A_n^{(e)}(X;Y)) + (M-1)P((\tilde{X}^n, \tilde{Y}^n) \in\in A_n^{(e)}(X;Y))\\ &= 0 + 2^{nR} \cdot 2^{-n(I(X;Y)-3\epsilon)}\\ &= 0 + 0\end{aligned}$$

by fixing ϵ such that $nR \le n(I(X;Y) - 3\epsilon)$.
where we used: $P((X^n(j), Y^n) \notin A_n^{(e)}(X;Y)|J = j) = P((X^n, Y^n) \notin A_n^{(e)}(X;Y))$ as well as $P((X^n(j), Y^n) \notin A_n^{(e)}(X;Y)|J = j) + P((X^n(k), Y^n) \in A_n^{(e)}(X;Y)|J = j)$
Therefore $\exists\{C_n\}$ a sequence of codebooks (at least one possible realization of our codebook) for which the probability of error will be less than the average probability of error for all the codebooks that could have been realized. Therefore, take at least one C_n with probability of error at least as good as $\mathbb{E}[P_e(C_n)]$, so we have a sequence of codebooks C_n with rate $\frac{1}{n}\log M_n \ge R$ and $P_e(C_n) \to 0$ since we choose C_n with lower probability of error than the expectation over all C_n. □

11 February 20

P_e: probability of error in decoding a message. This is an averaging over all the possible messages that might have been transmitted. Compare this with P_{max}: the maximum probability of error across any set of messages.

For a particular codebook $C^{(n)}$: $P_e = P(J \neq \hat{J}) = \sum_{j=1}^{M} P(J \neq \hat{J}|J = j)P(J = j)$
$P_{max} := \max_{i \le j \le m} P(J \neq \hat{J}|J = j)$

Claim: given C_n, there exists C'_n such that $|C'_n| \ge \frac{1}{2}|C_n|$ and $P_{max}(C'_n) \le 2P_e(C_n)$ (so P_{max} also vanishes).
Proof: Construct C'_n by removing from C_n the $|C_n|/2$ codewords with largest probability of coding error: $P(\hat{J} \neq J|J = j)$. The remaining codewords must each have $P(J \neq \hat{J}|J = j) \le 2P_e(C_n)$.

Therefore, $P_{max}(C'_n) \le 2P_e(C_n)$ and $|C'_n| = \frac{1}{2}|C_n|$. Combining with our direct part, if $R < I(X;Y)$, then exists sequence of codebooks with vanishing error probability. Further by this theorem, $\exists\{C'_n\}$ with $P_{max}(C'_n) \to 0$ as $n \to 0$ and $R = \frac{1}{n}\log|C'_n| = \frac{1}{n}\log|C_n|/2 \ge R - \frac{1}{n} \to R$ as $n \to \infty$: (dividing $|C_n|$ by 2 effectively reduces the rate by $\frac{1}{n}$.

Compression: $U_1, ..., U_n$ source components with $U_i \sim U$ i.i.d. Compression problem is $U_1, ..., U_n$ encoding it into n-bits → decoder → $U_1, ..., U_n$. [loseless compression].

More generally, we may relax loseless compression to lossy compression so that the output $V_1, ..., V_n$ to equal $U_1, ..., U_n$.

Recall rate = $\frac{n}{N}$ or the number of bits per source symbol. The smaller the rate the better (fewer bits required to represent source symbol). We also care about how far decoded sequence is from the input sequence.

Definition 20 (Distortion). *Given a distortion function: $d : \mathcal{U} \to \mathcal{V} \to [0, \infty)$, the distortion between U^N and V^N is $d(U^N, V^N) = \frac{1}{N}\sum_{i=1}^{N} d(u_i, v_i)$*

Definition 21 (Achievable). *A pair (r, D) is achievable if $\forall\epsilon > 0$, $\exists$ $(N, n$, encoder, decoder$)$ such that $\frac{n}{N} \le r + \epsilon$ and $\mathbb{E}[d(U^n, V^n)] \le D + \epsilon$. No more than $r + \epsilon$ bits per source symbol with expected distortion less than $D + \epsilon$.*

Definition 22 (Rate distortion function). *The rate-distortion function is $R(D) := \inf\{r|(r, D)$ is achievable.$\}$ The smallest rate with distortion approaching D.*

If $\mathcal{U} = \mathcal{V}$ [reconstruction alphabet is equal to the source alphabet], then "Hamming distortion" is $d(u, v) = \begin{cases} 0 \text{ if u = v} \\ 1 \text{ otherwise} \end{cases}$

If $\mathcal{U} = \mathcal{V} = \mathbb{R}$ then "squared error distortion" $d(u, v) = (u - v)^2$.

Definition 23. *The information rate-distortion function $R^{(I)}(D) := \min I(U; V)$ under all distributions where $\mathbb{E}[d(U, V)] \leq D$. U is given – it's the variable that characterizes our source distribution. Therefore for optimizing a joint distribution, we can only optimize over the conditional distribution of V given U: minimize over $P_{U|V}|\mathbb{E}[d(U, V)] \leq D$.*

Theorem 14. *$R(D)$ [the minimum rate that you can get away with by optimizing over all possible schemes in the world (i.e. encoding, decoding, etc.)] $= R^{(I)}(D)$ [a finite dimensional optimization problem by minimizing over $P_{U|V}$].*

Proof. Sketch of $R(D)$... $D_{max} = \min_V \mathbb{E}[d(U, V)] = 0$ [0 bits needed]. What value of V minimizes $\mathbb{E}[(U - v)^2]$ under squared error distortion? Simply set $V = \mathbb{E}[U]$ so we have $\mathbb{E}[(U - v)^2] = Var(U)$ so $D = Var(U) = D_{max}$ will have $R(D) = 0$. All D after this value will just set $D = Var(U)$ and therefore achieve $R(D) = 0$. □

Claim: $R(D)$ is a convex function of D. i.e. $\forall 0 < \alpha < 1$, D_0 and D_1: $R(\alpha D_0 + (1 - \alpha)D_1) \leq \alpha R(D_0) + (1 - \alpha)R(D_1)$.

Proof. Consider "time-sharing" compressor (i.e. scheme) employing (1) a good scheme (lossy decompressor) on the first $n \cdot \alpha$ source components for distortion level D_0. (2) use a good code (or scheme) for the remaining $n(1 - \alpha)$ source components for distortion level D_1.

It's overall distortion will be $\approx \alpha D_0 + (1 - \alpha)D_1$. It's overall rate will be $\approx$ (good scheme = comes close to the optimal–minimum number of bits per source symbol required) $R(D_0)$ bits per source symbol when encoding $n\alpha$ source symbols and $R(D_1)$ bits per source symbol when encoding $(1 - \alpha)n$ source symbols for D_1. Therefore the total rate will be: $\approx \frac{n\alpha R(D_0) + n(1-\alpha)R(D_1)}{n} = \alpha R(D_0) + (1 - \alpha)R(D_1)$.

We can't do better than $\alpha R(D_0)$ on the first $n\alpha$ source symbols and $(1 - \alpha)R(D_1)$ on the second $(1 - \alpha)n$ source symbols for rate given specified distortion amount. □

HW: Show that $R^{(I)}(D)$ is convex (without relying on the main result).
Example I: Let $U \sim Ber(p)$, $0 < p \leq \frac{1}{2}$ with Hamming distortion. Here $\mathcal{U} = \mathcal{V} = \{0, 1\}$.
Claim: $R(D) = \begin{cases} h_2(p) - h_2(D) \text{ for } 0 \leq D \leq p \\ 0 \text{ otherwise.} \end{cases}$

Proof. $R(D) = \min_{\mathbb{E}[d(U,V)] \leq D} I(U; V)$.

Assume U, V s.t. $\mathbb{E}[d(U, V)] =$ hamming distortion between $U, V = P(U \neq V) \leq D$. Recall our constraint $0 \leq D \leq p$ (trivially when $D \geq p$, $R(D) = 0$) so $P(U \neq V) \leq D \leq P \leq \frac{1}{2}$.

Then $I(U, V) = H(U) - H(U|V) = H(U) - H(U \bigoplus_2 V|V) \geq H(U) - H(U \bigoplus_2 V) = H(U) - h_2(P(U \neq V)) \geq H(U) - h_2(D) = h_2(p) - h_2(D)$ because $H(U \bigoplus_2 V)$ is binary and $h_2(D) \geq h_2(P(U \neq V))$.

To show equality, need to establish existence of a pair in the feasible set where inequalities turn into equalities. Specifically: $H(U \bigoplus_2 V|V) = H(U \bigoplus_2 V)$ as well as $h_2(P(U \neq V)) = h_2(D)$.

Equality if (1) we can find U, V such that $U \bigoplus_2 V$ independent of V and (2) $P(U \neq V) = D$.

Looking for V such that $U \bigoplus_2 V \sim Ber(D)$. Therefore $V \sim Ber(q)$. □

Exercise: Verify that $q = \frac{P-D}{1-2D}$ then $0 \leq q \leq \frac{1}{2}$ and that $U \sim Ber(p)$ when $V \sim Ber(q)$ (given $U = V \bigoplus_2 (U \bigoplus_2) V$)

12 February 22

Midterm will have "protector" status: count only if you do better on the midterm than the final. Otherwise, the final will get all the weight.

12.1 Lossy Compression

$U_1, U_2, ..., U_N$ i.i.d. $\sim U$. → Encoder that encodes this into n bits (i.e. output the index $J \in \{1, 2, ..., M\}$ where $M = 2^n$) → decoder → $V_1, ..., V_N$. $= V^N(J)$. Decoder takes an index into a reconstruction.

decoder $\iff$ $C_n = V^N(1), V^N(2), ..., V^N(M)$ [set of M possible reconstructions]. Basically, encode N possible indexes (i.e. which message) as binary $\log N$ bits (can loselessly represent N possible values).

Rate is bits per source symbol (i.e. U_i) $= \frac{n}{N} = \frac{\log M}{N}$ bits per source symbol (need $\log M$ bits for each source symbol U_i because it can take on 2^M different values).

Distortion $d(U^N, V^N) = \mathbb{E}[d(U^N, V^N)] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^N d(U_i, V_i)]$

$R(D)$ is the minimal rate needed to achieve distortion $\leq D$.

Main Result: $R(D) = \min_{\mathbb{E}[d(U,V)] \leq D} I(U; V) = R^{(I)}(D)$.

12.2 Continuous Lossy Compression

Compression of the Gaussian Source: U_i i.i.d. $\sim N(0, \sigma^2)$ under squared error distortion: $d(u, v) = (u - v)^2$.

Schemes at rate of $1 \frac{\text{bit}}{\text{source symbol}}$.
Scheme I: U_i → encoder → $B_i \in \{0, 1\}$ [1 bit] → decoder → V_i

What if we let $B_i = 1_{U_i \geq 0}$ [let the bit be the indicator for whether U_i is positive or not]. What would be the optimal reconstruction? V_i should be the mean of the positive half or the negative half. The optimal estimator of the squared error (distortion) is the mean: $V_i = \mathbb{E}[U_i|B_i]$.

$V_i(1) = \mathbb{E}[U_i|B_i = 1] = \mathbb{E}[U_i|U_i \geq 0]$ [V_i of the input argument J] $= \sqrt{\frac{2}{\pi}}\sigma$.

What is the distortion of this scheme? $\mathbb{E}[(U_i - V_i)^2] = \mathbb{E}[(U_i - V_i)^2|B_i = 1] = \mathbb{E}[U_i^2|U_i \geq 0] - (\mathbb{E}[U_i|U_i \geq 0])^2 = \sigma^2(1 - \frac{2}{\pi}) = 0.363\sigma^2$. Where we used the second moment squared minus the square of the first moment.

Scheme 2: Use 2 bits to describe 2 source symbols: U_1, U_2 → Encoder → 2 bits: $J \in \{1, 2, 3, 4\}$ → decoder → V_1, V_2. Can use 1 bit to tell if U_1 is greater than or less than 0. Set 2 bits to the horizontal and vertical axes (so one bit per source symbol detects whether this source symbol is ±).

Let's compute $R(D)$: assume $U \sim N(0, \sigma^2)$ and $\mathbb{E}[(U - V)^2] \leq D \leq \sigma^2$ [need the last inequality so the final mutual information is positive], then:

$$\begin{aligned}
I(U; V) &= h(U) - h(U|V) \\
&= h(U) - h(U - V|V) \\
&\geq h(U) - h(U - V) \\
&\geq h(U) - \frac{1}{2}\log 2\pi eD \\
&= \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi eD \\
&= \frac{1}{2}\log \frac{\sigma^2}{D}
\end{aligned}$$

$h(U - V|V)$ is simply the differential entropy of $U|V$ minus a constant (i.e. $V|V \Rightarrow$ invariant to shifts by a constant). $h(U - V)$ is an object with second moment less than D [$\mathbb{E}[(U - V)^2] \leq D$, so this is bounded above by the differential entropy of a Gaussian with second moment D.

Recall that we want to find $\min_{\mathbb{E}[(U-V)^2] \leq D} I(U; V) \geq \frac{1}{2}\log \frac{\sigma^2}{D}$ [lower bound the mutual information]. Looking at our series of inequalities, we seek a joint distribution under which:

1. $h(U - V|V) = h(U - V)$ [i.e. $U - V$ is independent of V].
2. $h(U - V) = \frac{1}{2}\log 2\pi eD$ [i.e. $U - V \sim N(0, D)$].

$U - V = N$, so V is independent of $N \sim N(0, D)$: $V \bigoplus N \to U \sim N(0, \sigma^2)$. Is there a RV V such that when you add to it and independent Gaussian N will yield the distribution of the source $U \sim N(0, \sigma^2)$. Yes! Simply $V \sim N(0, \sigma^2 - D)$. Then $R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}$ when $D \leq \sigma^2$ and 0 when $D > \sigma^2$.

$D(R)$ is the "distortion rate" function (inversion of $R(D)$) which is the minimal distortion achievable with rate $\leq R$.

$$\begin{aligned} D(R) = \frac{1}{2} \log \frac{\sigma^2}{D} &= R \\ \log \frac{\sigma^2}{D} &= 2R \\ \frac{\sigma^2}{D} &= 2^{2R} \\ D = \sigma^2 2^{-2R} & \end{aligned}$$

distortion as a function of rate R. Can get arbitrarily close to 0 distortion by sending more bits.

Compare: $D(1) = \frac{1}{4}\sigma^2$ is the best we can do across all schemes in the world working with arbitrarily big chunks of data. Compare this with what we're able to achieve with per-symbol quantization which is $0.363\sigma^2$.

By the law or large numbers (with high probability):

$$\sqrt{\sum_{i=1}^{N} U_i^2} \leq \sqrt{N\sigma^2}$$

Specifically $\|U^N - V^N\|_2 \leq \sqrt{ND}$.

To achieve distortion D, we need M, the size of the codebook, $\geq \frac{Vol(bigball)}{Vol(smallball)} = \frac{K_N \cdot (N\sigma^2)^{N/2}}{K_N \cdot (ND)^{N/2}} = \frac{\sigma^2}{D}^{\frac{N}{2}}$. Therefore the rate is $\frac{\log M}{N} \geq \frac{1}{2} \log \frac{\sigma^2}{D}$

13 February 27

13.1 Method of Types

$x^n = (x_1, x_2, ..., x_n)$, $x_i \in \mathcal{X} = \{1, 2, ..., r\}$ and let $N(a|x^n)$ (the number of times a occurs in the sequence $(x_1, ..., x_n)$:

$$N(a|x^n) = \sum_{i=1}^{n} 1_{\{X_i = a\}}$$

and

$$P_{x^n}(a) = \frac{N(a|x^n)}{n}$$

as the fraction of occurrences of a in the sequence $(x_1, ..., x_n)$.

Definition 24. *The empirical distribution of x^n is the probability vector (probability mass function represented by a vector of each symbol) $(P_{x^n}(1), P_{x^n}(2), ..., P_{x^n}(r))$ gives every element in the alphabet probability proportional to its occurrence in the sequence.*

$\mathcal{P}_n$ denotes the collection of all empirical distributions of sequences of length n. For any $P \in \mathcal{P}_n$, the **type** of P is $T(P) := \{x^n | P_{x^n} = P\}$.

Similarly, we can talk about the type of a sequence x^n, the type of the sequence x^n is $T_{x^n} = T(P_{x^n}) = \{\tilde{x}^n | P_{\tilde{x}^n} = P_{x^n}\}$ [the set of all sequences whose empirical distribution is equal to the distribution of the input sequence].

Example 6. *If $\mathcal{X} = \{0, 1\}$, then $\mathcal{P}_n = \{(1, 0), (\frac{n-1}{n}, \frac{1}{n}), ..., (0, 1)\}$*

$(1, 0)$ is the probability vector for the sequence of zeros: $(0, ..., 0)$, $(\frac{n-1}{n}, \frac{1}{n})$ is the probability vector for the sequence of 1 one and $n - 1$ 0s.

Example 7. *If $\mathcal{X} = \{a, b, c\}$, $n = 5$ and $x^n = (a, a, c, b, a)$, then $P_{x^n} = (\frac{3}{5}, \frac{1}{5}, \frac{1}{5})$.*

$T_{x^n} = \{(a, a, a, b, c), (a, a, b, a, c), ...\}$ with $|T_{x^n}| = \binom{5}{1} \cdot \binom{4}{1} = 20$ (5 spot to place b and then 4 spots to place c).

Theorem 15. $|\mathcal{P}_n| \leq (n+1)^{r-1}$

Proof. P_{x^n} is determined by $(N(1|x^n), N(2|x^n), ..., N(r-1|x^n))$ [r^{th} value is deterministic based on previous $r-1$ values] and $0 \leq N(i|x^n) \leq n$ [number of occurrences of any symbol is between 1 and n for all $i \in \mathcal{X}$].

Note – this is not a tight – does not take into account that all of the values in the sequence must sum up to n. Uses that each term $N(i|x^n) \leq (n+1)$. □

More Notation: for PMF $Q = (Q(1), ..., Q(r))$, write $H(Q)$ for $H(X)$ when $X \sim Q$. Similarly, $Q(x^n) = \prod_{i=1}^n Q(x_i)$ (iid source). For $S \subseteq \mathcal{X}^n$ [S is a collection of n-tuples], $Q(S)$ is the collective probability under an iid source Q, $Q(S) = \sum_{x^n \in S} Q(x^n)$

Theorem 16. $Q(x^n) = 2^{-n[H(P_{x^n}) + D(P_{x^n}||Q)]}$ $\forall x^n$ *[this is precise!]*

Proof.

$$\begin{aligned}
Q(x^n) &= \prod_{i=1}^{n} Q(x_i) \\
&= 2^{\log \prod_{i=1}^n Q(x_i)} \\
&= 2^{\sum i=1^n \log Q(x_i)} \\
&= 2^{\sum_{a \in \mathcal{X}} N(a|x^n) \log Q(a)} \\
&= 2^{n \sum_{a \in \mathcal{X}} \frac{N(a|x^n)}{n} \log Q(a)} \\
&= 2^{-n \sum_{a \in \mathcal{X}} P_{x^n}(a) \log \frac{1}{Q(a)} \frac{P_{x^n}(a)}{P_{x^n}(a)}} \\
&= 2^{-n[H(P_{x^n}) + D(P_{x^n}||Q)]}
\end{aligned}$$

Note that $H(P_{x^n})$ is the "empirical entropy". □

Theorem 17. *$\forall P \in \mathcal{P}_n$, $T(P)$ [set of all sequences whose empirical distribution is P]. How many sequences have an empirical distribution which is P? What is the size of $|T(P)|$?*

$\forall P \in \mathcal{P}_n$

$$\frac{1}{(n+1)^{r-1}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$$

Proof. Proof of the upper bound:

$$\begin{aligned}
1 &\geq P(T(P)) \\
&= \sum_{x^n \in T(P)} P(x^n) \\
&= \sum_{x^n \in T(P)} 2^{-n[H(P_{x^n}) + D(P_{x^n}||P)]} \\
&\text{note:} D(P_{x^n}||P) = 0 \text{ since summing over } x^n \in T(P) \\
&\text{note: } H(P_{x^n}) = H(P) \\
&= \sum_{x^n \in T(P)} 2^{-nH(P)} \\
&= |T(P)| \cdot 2^{-nH(P)}
\end{aligned}$$

Proof of lower bound:

Lemma 3. *For non-negative integers m, n, $\frac{m!}{n!} \geq n^{m-n}$.*

Proof. If $m \geq n$, then $\frac{m!}{n!} = m \cdot (m-1) \cdot ... \cdot (n+1) \geq n^{m-n}$ as there are $m - n$ factors in this product.
If $m < n$, then $\frac{m!}{n!} = \frac{1}{n \cdot (n-1) \cdot ... \cdot (m+1)}$ has $n - m$ factors; each of which is $\leq n$ so $\frac{1}{n \cdot (n-1) \cdot ... \cdot (m+1)} \geq \frac{1}{n^{n-m}}$ □

Lemma 4. *Multinomial coefficient:* $\binom{n}{n_1 \ldots n_k} = \binom{n!}{\prod_{i=1}^k n_i!}$
$\forall P, Q \in \mathcal{P}_n$: $P(T(P)) \geq P(T(Q))$ [the most probable type is the one coming from the true distribution]. $nP(a)$ is the number of times I see an "a" in the sequence whose empirical distribution is P.

$$\begin{aligned}
\frac{P(T(P))}{P(T(Q))} &= \frac{|T(P)| \cdot \prod_{a \in \mathcal{X}} p(a)^{n \cdot P(a)}}{|T(Q)| \cdot \prod_{a \in \mathcal{X}} p(a)^{n \cdot Q(a)}} \\
&= \frac{\binom{n}{n \cdot p(1) n \cdot p(2) \ldots n \cdot p(r)}}{\binom{n}{nQ(1) \ldots n \cdot Q(2)}} \cdot \prod_{a \in \mathcal{X}} p(a)^{n[P(a)-Q(a)]} \\
&= \prod_{a \in \mathcal{X}} \frac{(nQ(a))!}{(nP(a))!} p(a)^{n[P(a)-Q(a)]} \\
&\geq \prod_{a \in \mathcal{X}} [np(a)]^{nQ(a)-nP(a)} p(a)^{n[P(a)-Q(a)]} \\
&= \prod_{a \in \mathcal{X} | n^{n[Q(a)-P(a)]}} \\
&= n^{n \sum_{a \in \mathcal{X}} Q(a) - P(a)} \\
&= n^{n \cdot 0} = 1
\end{aligned}$$

Where $\sum_{a \in \mathcal{X}} Q(a) - P(a) = 0$ as both Q and P are pmfs.

Now we are ready to prove the lower-bound in full.

$$\begin{aligned}
1 &= P(\mathcal{X}^n) \\
&= \sum_{Q \in \mathcal{P}_n} P(T(Q)) \\
&\leq |\mathcal{P}_n| \cdot \max_{Q \in \mathcal{P}_n} P(T(Q)) \\
&= |\mathcal{P}_n| P(T(P)) \\
&= |\mathcal{P}_n| \cdot |T(P)| \cdot 2^{-n[H(P)+D(P||P)]} \\
&\leq (n+1)^{r-1} |T(P)| \cdot 2^{-nH(P)}
\end{aligned}$$

□

Note: by Theorem 2, we know that $\forall P \in \mathcal{P}_n$ and any source Q [not necessarily an empirical distribution], $Q(T(P))$ [the probability of getting a sequence whose empirical distribution is P] $Q(T(P)) = |T(P)| \cdot 2^{-n[H(P)+D(P||Q)]}$. By Theorem 3, $\frac{1}{(n+1)^{r-1}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$, we have

$$\frac{1}{(n+1)^{r-1}} 2^{-nD(P||Q)} \leq Q(T(P)) \leq 2^{-nD(P||Q)}$$

This is the bottom line–super important. The role of the true source is played by Q: data is generated iid from Q, but what is the probability that it looks like it came from source P: the answer is given in terms of the relative entropy $D(P||Q)$.

14 February 29

14.1 Types

For positive $\{\alpha_n\}, \{\beta_n\}$ we say $\alpha_n = \beta_n$ for $\frac{1}{n} \log \frac{\alpha_n}{\beta_n} \to 0$ as $n \to 0$

i.e. $\alpha_n := 2^{n\gamma} \iff \alpha_n = 2^{n(\gamma + \epsilon_n)}$ where $\epsilon_n \to 0$ as $n \to 0$ i.e. $\alpha_n = 2^{n\gamma} \cdot 2^{n\epsilon_n}$ so that the second term vanishes as $n \to 0$

Recap: $x^n \in \mathcal{X}^n$, $P_{x^n}(a) = \frac{N(a|x^n)}{n}$, and the empirical distribution of a sequence x^n is P_{x^n}.

We saw $\mathcal{P}_n$ is the set of all empirical distributions that can be induced by a sequence of length n. We've seen that $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} - 1$ (the number of different types is less than or equal to a function of n and $\mathcal{X}$).

We've also seen that $\forall$ pmf Q and any $P \in \mathcal{P}_n$: [the probability that I get a sequence of type P]. T(P): all the empirical distributions of P. The probability you'll see an empirical distribution from Q that belongs in T(P).

$$\frac{1}{(n+1)^{r-1}} 2^{-nD(P||Q)} \leq Q(T(P)) \leq 2^{-nD(P||Q)}$$

In particular $Q(T(P))$ is exponentially small when Q has distribution highly different from P: $Q(T(P)) = 2^{-nD(P||Q)}$ and moreover, $|T(P)| = 2^{nH(P)}$.

14.2 Strong Typicality

Definition 25. *A sequence in $x^n \in \mathcal{X}^n$ is strongly d-typical with respect to some pmf p if $|p_{x^n}(a) - p(a)| \leq \delta \cdot p(a)$ $\forall a \in \mathcal{X}$.*

Let $T_\delta(P)$ denote the set of all such sequences.

In HW, will show $T_\delta(P) \subseteq A_\epsilon(P)$ for $\epsilon = \delta \cdot H(P)$.
2) $Q(T_\delta(P)) = 2^{-n[D(P||Q) - \epsilon(\delta)]}$ where $\epsilon(\delta) \to 0$ as $d \to 0$ is determined by the most probable type that comprises the strongly typical set. Allowing slack in the empirical distribution: allowing sequences that are slightly different from P. That's where the ϵ is coming from.

Definition 26. *For $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$ their joint empirical distribution is $P_{x^n,y^n} = \frac{1}{n}\sum_{i=1}^n 1_{x_i=x, y_i=y}$.*
(x^n, y^n) is strongly jointly δ-typical w.r.t. $P_{X,Y}$ if $|P_{x^n,y^n}(x,y) - P_{X,Y}(x,y)| \leq \delta P_{X,Y}(x,y) \forall x \in \mathcal{X}, y \in \mathcal{Y}$.

Let $T_\delta(P_{X,Y}) = T_\delta(X,Y)$ denote the set of all such pairs (x^n, y^n).

HW 3): $(x^n, y^n) \in T_\delta(X,Y) \Rightarrow d(x^n, y^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, y_i) \leq (1+\delta)\mathbb{E}[d(X,Y)]$

HW 4): X^n iid $\sim X$, $Y^n \sim Y$ iid, and X^n, Y^n are independent, then

$$Pr((X^n, Y^n) \in T_\delta(X,Y)) = 2^{-n[I(X;Y) - \epsilon(\delta)]}$$

where $\epsilon(\delta) \to 0$ as $\delta \to 0$. We can see this from $I(X;Y) = D(P_{X,Y}||P_X \times P_Y)$.

$U_1, U_2, ..., U_n$ iid $\sim U \to$ encoder $\to J \in \{1, ..., M\}$ [possible indices] $\to$ decoder takes as input a sequence of bits representing the index $\to V^N(J)$

Decoder $\iff$ c_N [codebook] $= \{V^N(j)\}_{j=1}^M$ is the possible reconstructions. Rate $= \frac{\log M}{N}$ bits per symbol. And $d(U^N, V^N)$ is the per-symbol distortion: $= \frac{1}{N}\sum_{i=1}^N d(U_i, V_i)$ and $d(U^N, c_N) = \min_{1 \leq j \leq N} d(U^N, V^N(j))$.

Given a codebook, the best scheme would be to give the index of that codeword in the codebook that is closest to the source sequence we are trying to represent.

Main result: $R(D) = \min_{\mathbb{E}[d(U,V)] \leq D} I(U;V) = R^{(I)}(D)$

Proof. Sketch of direct part: Fix U, V such that $\mathbb{E}[d(U,V)] \leq D$ and $\epsilon > 0$. Need to show that for sufficiently large N, $\exists C_N$ [codebook] such that $|C_N|$ (or equivalently, the rate) needs to be within ϵ of the mutual information between U, V: $|C_N| \leq 2^{n[I(U;V)+\epsilon]}$ and $\mathbb{E}[d(U^N, C_N)]$ [encode closest member to it] $\leq D + \epsilon$.

Give me any pair in the feasible set, and I can find you a scheme whose rate is within ϵ of the mutual information between them and whose distortion is within ϵ of D. [Direct part of the main result].

Generate $C_N = \{V^N(j)\}_{j=1}^M$ iid $\sim V$.

$P((U^N, V^N(j)) \in T_\delta(U,V)) \approx 2^{-nI(U;V)}$. If $M = 2^{n[I(U;V)+\epsilon]}$, then with overwhealming probability, will find a $V^N(j)$ that is jointly typical with U^N. This is good because the distortion between the source sequence between $U, V^N(j)$ will be similar to $\mathbb{E}[d(U,V)] \leq D$:

$$P(d(U^N, C_N) \tilde{<} D) \geq P(\cup_{j=1}^m (U^N, V^N(j)) \in T_\delta(U,V)) \approx 1 \Rightarrow \mathbb{E}[d(U^N, C_N)] \tilde{\leq} D$$
$$\Rightarrow \exists c_N \text{ such that } \mathbb{E}[d(U^N, c_N)] \tilde{\leq} D$$

to not exceed $D + \epsilon$ with $|c_N| = 2^{n(I+\epsilon)}$. In particular, having a reconstructed sequence in the codebook within distortion D of U. If we generate a codebook with sufficient size, then with very high probability, this codebook will have distortion wihtin the realized source less than D [because there will be a joint typical sequence]. In particular, we can extract one specific codebook c_N [a realization of codebooks].

Proof of the converse part: consider a scheme with $\mathbb{E}[d(U^N, V^N)] \leq D$. Then, let's consider what must be satisfied. Consider the entropy associated with the reconstruction:

$$\begin{aligned}
\log M \geq H(V^N) \geq I(U^N, V^N) &= H(U^N) - H(U^N|V^N) \\
&= \sum_{i=1}^{N} H(U_i) - H(U_i|U^{i-1}, V^N) \\
&\geq \sum_{i=1}^{N} H(U_i) - H(U_i|V_i) \\
&= \sum_{i=1}^{N} I(U_i; V_i) \\
&\text{a mutual information} \\
&\geq \sum_{i=1}^{N} R^{(I)}(\mathbb{E}[d(U_i, V_i)]) \\
&\text{the minimum mutual information you can get between U and V} \\
&\geq N \sum_{i=1}^{N} \frac{1}{N} R^{(I)}(\mathbb{E}[d(U_i, V_i)]) \\
&\geq N \sum_{i=1}^{N} \frac{1}{N} R^{(I)}(\mathbb{E}[d(U_i, V_i)]) \\
&\geq N R^{(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[d(U_i, V_i)])} \\
&\geq N R^{(\frac{1}{N} \mathbb{E}[d(U^N, V^N)])} \\
&\text{However, } \mathbb{E}[d(U^N, V^N)] \leq D \\
&\geq N \cdot R^{(I)}(D) \\
&\text{since D upper bounds } \mathbb{E}[d(U^N, V^N)]
\end{aligned}$$

Therefore the rate of this scheme is $\frac{\log M}{N} \geq R^{(I)}(D)$. An arbitrary scheme that satisfies the distortion constraint necessarily has rate at least the minimum over all mutual informations: $R^{(I)}(D)$. □

15 March 5

Reminders:

1. From proof of converse in reliable communication: $X^n \rightarrow$ Channel $\rightarrow Y^n$ [noisy components] $I(X^n, Y^n) \leq n \cdot C$ [mutual information is bounded above by n times the channel capacity].
2. From proof of converse to lossy compression: if $\mathbb{E}[d(U^N, V^N)] \leq D$ then $I(U^n, V^n) \geq NR(D)$ where $R(D)$ is the minimum mutual information.
3. $U - V - W - Z$ [Markov chain] $\Rightarrow I(V, W) \geq I(U, Z)$

Joint Source Channel Coding (JSCC)
Have N source components $U_1, ..., U_N$ with each U_i drawn i.i.d. from a source U.

$U_1, ..., U_N \rightarrow$ Encoder/Transmitter $\rightarrow X_1, ..., X_n \rightarrow$ [Memoryless Channel $P_{Y|X}$] $\rightarrow Y_1, ..., Y_n \rightarrow$ Receiver/Decoder $\rightarrow$ the reconstruction signal $V_1, ..., V_N$.

We can about the rate: N bits for n channel uses [number of source symbols that I communicate per channel use]. We want the rate to be high: we communicate many bits per a single channel usage.

We also have the expected distortion of the scheme: $\mathbb{E}[d(U^N, V^N)]$ [average per-symbol distortion] measured by a component-wise distortion criteria.

We have $U^N - X^N - Y^N - V^N$ has the Markov relation: noise gets added from X^N to Y^N: knowing U^N does not change $P(Y^N|X^N)$.

Definition 27. *A pair (ρ, D) for [rate, distortion] is achievable if $\forall \epsilon > 0$ $\exists$ a scheme (i.e. N, n, encoder, decoder [how many source symbols you'll be encoding, channel uses, encoder, decoder]) such that $\frac{N}{n} \geq \rho - \epsilon$ and $\mathbb{E}[d(U^N, V^N)] \leq D + \epsilon$.*

Note: For any scheme with $\mathbb{E}[d(U^N, V^N)] \leq D$, then $N \cdot R(D) \leq I(U^N; V^N)$ [from Reminder #2]. And further, due to the Markov chain,

$$N \cdot R(D) \underbrace{\leq}_{\text{from \#2}} I(U^N; V^N) \underbrace{\leq}_{\text{from \#3}} I(X^n; Y^n) \underbrace{\leq}_{\text{from \#1}} n \cdot C$$

Therefore, $\frac{N}{n} \leq \frac{C}{R(D)}$. For any scheme in the world, if it achieves end-to-end distortion less than D, then necessarily its rate is less than $\frac{C}{R(D)}$.

Conclusion I: if (ρ, D) is achievable, then necessarily, $\rho \leq \frac{C}{R(D)}$.

Take N large. Consider a "separation" scheme based on comprising:

1. a good rate distortion code for N source symbols at distortion D. The minimum number of bits per source symbol that we need $R(D)$ for a specific distortion D. So we will need $\approx N \cdot R(D)$ bits.
2. a good channel coding scheme for reliably communicating $N \cdot R(D)$ bits.

How many channel uses will suffice? We want to communicate $N \cdot R(D)$ bits through a channel that has a certain capacity C [communicating C bits per channel use]. Then we need $\approx \frac{N \cdot R(D)}{C}$ channel uses will suffice.

$\Rightarrow$ as long as n is sufficiently large (i.e. $n \geq \frac{N \cdot R(D)}{C}$), then $\frac{N}{n} = \frac{C}{R(D)}$ and your scheme will work. In other words, you will get expected distortion no more than D.

Conclusion II: If $\rho \leq \frac{C}{R(D)}$, then (ρ, D) is achievable.

Putting together **Conclusion 1** and **Conclusion 2**, we have a complete view of what is achievable and what is not.

Theorem 18 (JSCC Separation Theorem: Fundamental Result of this Course)**.** *(ρ, D) is achievable if and only if $\rho \leq \frac{C}{R(D)}$*

Note that the bits here have popped out of "thin air" as part of the solution. There is no stipulation about working in bits. What we have found is that if we want to achieve optimal performance, the right architecture is to represent it in bits and then digitize it and then protect those bits from noise/error. Separate communities that worry about compression, others that worry about error correction, but they're really combined together.

Bits are not the only architecture that gets us optimal performance; however, they are guaranteed to help us achieve optimal performance.

Example I: Binary Source + Binary Channel
$U \sim Ber(p)$ with $0 \leq p \leq \frac{1}{2}$, $BSC(\delta)$ with $0 \leq \delta \leq \frac{1}{2}$ under a Hamming distortion function.

Recall $R(D) = \begin{cases} h_2(p) - h_2(D) \text{ for } 0 \leq D \leq P \\ 0 \text{ for } D > P \end{cases}$

Recall $C = 1 - h_2(\delta)$

From JSCC theorem: (ρ, D) is achievable iff $\rho \leq \frac{1-h_2(\delta)}{h_2(P)-h_2(D)}$. Let's now check that the edge cases are consistent with what we know from the vanilla communication problem and rate distortion problem.

Sanity Check I: Reliable communication setting: $p = \frac{1}{2}$, $D = 0$, we get $\rho \leq \frac{1-h_2(\delta)}{h_2(p)-0} = \frac{1-h_2(\delta)}{1}$

Sanity Check II: Lossy compression: $\delta = 0$ [cross-over probability is 0]. ρ is achievable iff $\rho \leq \frac{1-0}{h_2(P)-h_2(D)}$ [in the context of compression, the rate is inverted]. Rate [in the context of compression] is the ratio between number of bits we need (i.e. number of channel uses in a "clean channel") divided by the number of source symbols. $= \frac{\#\text{ bits } = n}{N} \geq h_2(P) - h_2(D)$.

Interesting special case: $p = \frac{1}{2}$, $D = \delta$ [distortion is equal to the channel cross-over probability in BSC(δ)]. In this case, connect the source straight to the channel.

Consider the scheme: $X_i = U_i$ [what you're putting into the channel is the ith bit and V_i is the trivial decoding $V_i = Y_i$ since the distortion is δ. We're ok with δ bits flipping.

This scheme achieves distortion δ [channel cross-over probability] and the rate [back to the rate definition from the communication context] is 1. How does this compare to the optimal for this setting? The optimal rate is $\frac{C}{R(D)} = \frac{1-h_2(\delta)}{1-h_2(D)} = \frac{1-h_2(D)}{1-h_2(D)} = 1$.

In this very special case, the trivial scheme achieves the optimal.

Example II: $U \sim N(0, \sigma_S^2)$ **and** $d(u,v) = (u-v)^2$
We use the AWGN channel: $Y_i = X_i + N_i$ where $N_i \sim N(0, \sigma_N^2)$. We also have power constraint: P.

Recall $R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma_S^2}{D} \leq D \leq \sigma_S^2 \\ 0 \text{ if} D > \sigma^2 \end{cases}$ where S denotes the variance of the source and N denotes the variance of the additive noise.
$C(P) = \frac{1}{2}\log(1 + \frac{P}{\sigma_N^2}$.

BY JSCC "separation" theorem: (ρ, D) is achievable iff $\rho \leq \frac{\log(1+\frac{P}{\sigma_N^2})}{\log\frac{\sigma_S^2}{D}}$.

Question: what is the best distortion for a fixed rate? Previously fixed distortion and found the best rate. What is the best (smallest) distortion at rate 1 source symbol per channel use?

It's going to be the D so that $\frac{\log(1+\frac{P}{\sigma_N^2})}{\log\frac{\sigma_S^2}{D}} = 1$

$$\begin{aligned} \log\frac{\sigma_S^2}{D} &= \log(1 + \frac{P}{\sigma_N^2}) \\ \frac{\sigma_S^2}{D} &= 1 + \frac{P}{\sigma_N^2} \\ D &= \frac{\sigma_N^2 + P}{\sigma_N^2\sigma_S^2} \end{aligned}$$

16 March 12

16.1 General Communication Setting (i.e. JSCC)

For source U_i i.i.d. $\sim U$ $U_1, ..., U_N \to$ Encoder $\to X^n \to$ Noisy Channel $\to Y^n \to$ Decoder $\to V^N$.

Rate is $\frac{N}{n}$ source symbols per channel use (send a single X through the channel at a time).

Expected distortion: $\mathbb{E}[d(U^N, V^n)]$. Then (ρ, D) is achievable if ...

Theorem 19. *Separation Theorem: (ρ, D) is achievable $\iff$ $\rho \leq \frac{C}{R(D)}$ [ρ in $\frac{N}{n}$].*

Example: $U_i \sim N(0, \sigma_S^2)$ (σ_s is the std of the source). And the channel is the AWGN channel: $Y_i = X_i + N_i$ where $N_i \sim$ i.i.d. $N(0, \sigma_N^2)$ under transmission power constraint P. Recall $R(D) = \frac{1}{2}\log\frac{\sigma_S^2}{D}$ for $0 < D \leq \sigma^2$ and capacity for a given power constraint $C(P) = \frac{1}{2}\log(1 + \frac{P}{\sigma_N^2})$. Assume squared error distortion.

Separation theorem: (ρ, D) is achievable $\iff$ $\rho \leq \frac{1/2\log(1+\frac{P}{\sigma_N^2})}{1/2\log(\frac{\sigma_S^2}{D})}$. A best way is to digitize your source with bits and then protect these bits from disruption.

Question: For rate = 1 source symbol per channel use, what is the best (or smallest) achievable distortion?
Answer: the D that solves $1 = \frac{1/2\log(1+\frac{P}{\sigma_N^2})}{1/2\log(\frac{\sigma_S^2}{D})}$.

$$1 = \frac{\log(1+\frac{P}{\sigma_N^2})}{\log(\frac{\sigma_S^2}{D})}$$
$$D = \frac{\sigma_N^2\sigma_S^2}{\sigma_N^2 + P}$$

Consider: U_i and stick them straight into the channel. $X_i = U_i$ but we have a power constraint to satisfy, so $X_i = \sqrt{\frac{P}{\sigma_S^2}}U_i$ so that $\mathbb{E}[X_i^2] = P$ [satisfies the power constraint]. And then let the decoding be $V_i = \mathbb{E}[U_i|Y_i]$. U_i gaussian, noise is gaussian, etc. so conditional expectation is a linear function and $\mathbb{E}[U_i|Y_i] = c\cdot Y_i$ where c is a constant $c = \rho\cdot\frac{Var(U_i)}{Var(Y_i)}$ with $\rho = \frac{\mathbb{E}[(U_i\cdot Y_i)]}{\sqrt{Var(U_i)\cdot Var(Y_i)}}$. And moreover, $\mathbb{E}[(U_i - V_i)^2] = (1-p^2)\cdot Var(U_1) = \frac{\sigma_N^2\sigma_S^2}{p+\sigma_N^2}$.

Moral of the story: this simple scheme is actually optimal because the expected distortion is equal to the lowest possible distortion via the JSCC theorem. Therefore, we don't need to search for a better scheme.

16.2 (Near) Lossless Compression

X_i iid $\sim X$. Have $X^n \rightarrow$ encoder $\rightarrow$ m bits $\rightarrow$ decoder $\rightarrow \hat{X}^n$. Then a rate R is achievable if $\forall\epsilon > 0$ there exists a scheme (encoder, decoder, n, m) such that $\frac{m}{n} \leq R + \epsilon$ and $P(\hat{X}^n \neq X^n) \leq \epsilon$ [setting of near lossless compression].

And recall that the best we can do is $nH(X)$ bits. This is very simple, fixed-block (i.e. fixed length) lossless compression. Variable length would imply that n can change...
Main Theorem: R is achievable iff $R \geq H(X)$.

16.3 (near) Lossless Compression with Side Information

Still want to communicate X_i, but now correlated with Y_i: (X_i, Y_i) iid $\sim (X, Y)$.

X_i iid $\sim X$. Have $X^n \rightarrow$ encoder $\rightarrow$ m bits $\rightarrow$ decoder $\rightarrow \hat{X}^n$

But now Y^n is available to both the encoder and decoder. We call Y^n the side-information sequence. If X^n is your genome, then Y^n is a publicly available reference genome.

Theorem 20. *R is achievable* $\iff R \leq H(X|Y)$.

Conditionally typical set has size $2^{nH(X|Y)}$. Therefore, only need the bits to index the elements in the conditionally independent set. Alternatively, for every possible value of y, can look at $H(X|Y = y)$ and use the appropriate scheme and then average across all possible values of y (based on $p(y)$). **If we gave this as a homework exercise, then you could establish $R \geq H(X|Y)$ completely and formally**. Use AEP or argue for describing an encoding scheme for each value of $Y = y$ and then averaging across these by the probability of $p(y)$.

Can transfer all the theorems from vanilla loseless compression to the conditional case.

Different settings:
Side Information Sequence is known only at the Encoder: If only the encoder has access to Y^n then the limit on compression is still $R \geq H(X)$. In this case, it's just extra noise for the encoder. Not helpful at all.

Side Information Sequence is known only at the Decoder: If only the decoder has access to Y^n. This is called (near) lossless compression with decoder side information.

Theorem 21 (Slepian-Wolf 1973)**.** *R is achievable iff* $R \geq H(X|Y)$.

In this setting, you can do essentially as well as if Y^n is available to the encoder. **Example** $X \sim Ber(0.5)$, $Y \sim Ber(0.5)$, $P(X \neq Y) = \delta$.

$X \rightarrow \text{BSC}(\delta) \rightarrow Y$. For example X is your genome and Y is the reference genome. We did not explicitly define this relationship; nature did. Note $H(X|Y) = h_2(\delta)$ [simply the entropy of the noise that flips the values].

Consider the scheme: randomly assign each of the 2^n possible source sequences a color (or equivalently "bin") from a set of size $2^{n(h_2(\delta)+\epsilon)}$.

This color assignment is shared and known to both encoder and decoder.

Encoding: use $n(h_2(\delta)+\epsilon)$ bits to inform decoder of the color of X^n.
Decoding: $\hat{X}^n$ be the source sequence in the hamming ball of radius $n\delta$ around the side information sequence Y^n with this color which was described to the decoder by the encoder.

This is similar to random hashing from CS where the color (or "bin") is the random hash bin. We can also extend this setting to the distributed setting.

16.4 (near) lossless distributed compression

$X^n \rightarrow$ Encoder 1 $\rightarrow m_X$ bits $\rightarrow$
i.e. your genome.
$Y^n \rightarrow$ Encoder 2 $\rightarrow m_Y$ bits $\rightarrow$ i.e. other genome.

The cloud will process both jointly with a single decoder to loselessly reconstruct $\hat{X}^n$, $\hat{Y^n}$ is the more general setting. In this setting, what is the optimal tradeoff in the rates of the two compression schemes.

17 March 14

Theorem 22 (Slepian and Wolf 1973). *R is achievable $\iff R \geq H(X|Y)$*

Distributed (near) lossless compression:
X^n encoded as m_X bits and Y^n encoded as m_Y bits. Then they share a decoder that outputs $\hat{X}^N, \hat{Y}^N$.

We say a rate pair (R_X, R_Y) is achievable if $\forall \epsilon > 0$, $\exists n$ and a scheme (2 encoders + decoder) such that $\frac{m_X}{n} \leq R_X + \epsilon$ and $\frac{m_Y}{n} \leq R_Y + \epsilon$ and $P((\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)) \leq \epsilon$.

Theorem 23 (Slepian-Wolf 1973). *(R_X, R_Y) is achievable $\iff R_X \geq H(X|Y)$ and $R_Y \geq H(Y|X)$ and $R_X + R_Y \geq H(X, Y)$*

17.1 Distributed Lossy Compression [Multi-Terminal Source Coding]

$X^n \rightarrow$ encoder $\rightarrow m_X$ bits $\rightarrow$ shared decoder.
$Y^n \rightarrow$ encoder $\rightarrow m_Y$ bits $\rightarrow$ shared decoder.
Output of the shared decoder is $\hat{X}^n$, $\hat{Y}^n$. (R_X, R_Y, D_X, D_Y) is achievable if $\forall \epsilon > 0$...

For this 4-dimensional tuples, what is achievable and what is not achievable. Open question — able to give inner and outer bounds. Our understanding of this problem is to fix a pair of distortions (distortion level) and then talk about achievable rates.

Lossy compressor with decoder Side Information:
$X^n \rightarrow$ encoder $\rightarrow$ m bits $\rightarrow$ decoder (that takes in $Y^n \rightarrow \hat{X}^n$
$R_{WZ}(D)$ [WZ stands for Waner-Ziv]. In general $R_{X|Y}(D) \leq R_{WZ}(D)$ [in general strict]. When X, Y are Gaussian, and squared-error distortion, $R_{X|Y}(D) = R_{WZ}(D)$.
Multiple Access Channel (MAC) Message $J \in \{1, ..., M\}$ in a set of M possible messages communicated through a noisy channel:
$J \rightarrow$ encoder $\rightarrow X^n \rightarrow$ Noisy Channel $P_{Y|X}$ $\rightarrow Y^n \rightarrow$ decoder $\rightarrow \hat{X}$.
$J_2 \in \{1, ..., M_2\}$