EE229A Notes

Ben Mildenhall

November 15, 2016

September 1

Homework notes

Coin weighing problem: use entropy for the first part. Second part is just very hard trial and error (don't worry about it too much).

Getting random bits from a biased coin: use entropy.

A puzzle

Person A observes i.i.d. fair coin flips $X_1, \dots, X_n$ (for a large n). Person B observes $Y_1, \dots, Y_n$ from Bern(1/2) also, but they're not independent: $P(X_i = Y_i) = .98$. We want them to output Z_A and Z_B respectively (taking values in $\{0, 1\}$) so that $P(Z_A = 0) = 1/2$, same for Z_B. Can they agree with probability .99?

Connecting entropy to estimation

In estimation we have some X (nature) influencing an observation Y, and we want to use this to get an estimate $\hat{X}(Y)$.

A measure for goodness of estimator is

$$P_e = P(X \neq \hat{X}(Y)),$$

the probability of error. (Going forward we use $\hat{X} = \hat{X}(Y)$ as shorthand.)

Theorem: (Fano's Inequality)

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X \mid \hat{X}).$$

Proof: Define an indicator variable $E = 1\{X \neq \hat{X}\}$.

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(X \mid \hat{X}) + H(E \mid X, \hat{X}), \text{ but } H(E \mid X, \hat{X}) = 0, \\ &= H(E \mid \hat{X}) + H(X \mid E, \hat{X}), \text{ expand first expression with chain rule.} \end{aligned}$$

But $H(E \mid \hat{X}) \leq H(E) = H(P_e)$ and

$$H(X \mid E, \hat{X}) = \sum_{e=0,1} p_E(e) H(X \mid \hat{X}, E = e) = 0 + p_E(1) H(X \mid \hat{X}, X \neq \hat{X}) \leq P_e \log |\mathcal{X}|.$$

Corollary:

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

for $p_e = \min_{\hat{X}} P(\hat{X}(Y) \neq X)$.

Proof: $X \to Y \to \hat{X}(Y)$, so $I(X; \hat{X}) \leq I(X; Y)$. Thus

$$H(X) - H(X \mid \hat{X}) \leq H(X) - H(X \mid Y) \Rightarrow H(X \mid Y) \leq H(X \mid \hat{X}).$$

The corollary follows directly.

What it means to describe something with bits

We can think of a function as $f : \mathcal{X} \to \{0,1\}^\star = \{\Lambda, 0, 1, 00, 01, 10, 11, 000, \ldots\}$. f should be one-to-one so that $x \neq x' \Rightarrow f(x) \neq f(x')$.

ℓ = length function. So $\ell(f(x)) = (\ell \circ f)(x)$ = length of description of x in bits.

The expected length of the description of x is $E[\ell \circ f(x)]$. We would like to bound this.

$$\begin{aligned}
E[\ell \circ f(X)] &= \sum_{x \in \mathcal{X}} p_X(x) \ell \circ f(x) - H(x) + H(x) \\
&= \sum_x p_X(x) \log 2^{\ell \circ f(x)} + \sum_x p_X(x) \log p_X(x) + H(x) \\
&= \sum_x p_X(x) \log \frac{p_X(x)}{2^{-\ell \circ f(x)}} + H(x) \\
&= \sum_x p_X(x) \log \frac{p_X(x)}{2^{-\ell \circ f(x)} / \sum_{x'} 2^{-\ell \circ f(x')}} + H(x) - \sum_x p_X(x) \log \sum_{x'} 2^{-\ell \circ f(x')} \\
&= \sum_x p_X(x) \log \frac{p_X(x)}{2^{-\ell \circ f(x)} / \sum_{x'} 2^{-\ell \circ f(x')}} + H(x) - \log \sum_x 2^{-\ell \circ f(x)}
\end{aligned}$$

First term is a KL divergence, so it is nonnegative and we get

$$E[\ell \circ f(X)] \geq H(x) - \log \sum_x 2^{-\ell \circ f(x)}.$$

Since $\sum_x 2^{-\ell \circ f(x)} \leq \log 2|\mathcal{X}|$, we get

$$E[\ell \circ f(X)] \geq H(x) - \log \log 2|\mathcal{X}|.$$

Now, for $X_1, \ldots, X_n \sim$ i.i.d. X, then the size of the support of $X^n = (X_1, \ldots, X_n)$ is $|\mathcal{X}|^n$. So

$$\frac{1}{n} E \ell \circ f_n(X^n) \geq H(x) - \frac{\log(1 + n \log |\mathcal{X}|)}{n}.$$

Thus the average number of bits needed to describe each outcome is at least $H(X) - O\left(\frac{\log n}{n}\right)$ when we take a sequence of n samples.

The A.E.P. (Asymptotic equipartition property)

Consider coin flips with $P(X = 0) = p$. Suppose $X^n = (X_1, \dots, X_n)$ are i.i.d. copies of X. The "typical" X^n has pn 0's and $(1-p)n$ 1's. What is the probability $p_{X^n}(x^n)$ of observing this exact sequence?

$$p_{X^n}(x^n) = p^{pn}(1-p)^{(1-p)n} = 2^{-nH(X)}.$$

Surprisingly, entropy emerges. So the probability of the typical sequence is exponentially small, with a coefficient controlled by the entropy.

From HW0, if $X_1, X_2, \dots$ are i.i.d. $\sim X$ and we have a function f, then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to Ef(X) \text{ in probability,}$$

which means that

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef(X)\right| \geq \epsilon\right) = 0$$

for all $\epsilon > 0$.

Apply this to the function $f(\cdot) = \log \frac{1}{p_X(\cdot)}$.

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n} \log\frac{1}{p_X(X_i)} = \frac{1}{n}\log\frac{1}{\prod_{i=1}^{n} X_i} = \frac{1}{n}\log\frac{1}{p_{X^n}(X^n)}$$

Thus we get that

$$\frac{1}{n}\log\frac{1}{p_{X^n}(X^n)} \to H(X)$$

in probability.

Stated another way,

$$p_{X^n}(X^n) \approx 2^{-nH(X)}$$

with high probability. So the probability of a random observed sequence is very likely to be close to this entropy approximation.

The A.E.P.: $-\frac{1}{n}\log p_{X^n}(X^n) \to H(X)$ in probability.

What is a "typical" observation?

Defn: For $\epsilon > 0$ and n, define the "typical set" $A_\epsilon^{(n)} \subseteq \mathcal{X}^n$ as

$$A_\epsilon^{(n)} = \{x^n : 2^{-n(H(x)+\epsilon)} \leq p_{X^n}(x^n) \leq 2^{-n(H(x)-\epsilon)}\}.$$

September 6

Recap

Some examples where entropy is important:

1. If $f : \mathcal{X} \to \{0,1\}^*$ is one-to-one, then $E\ell \circ f(X) \geq H(X) - \log\log 2|\mathcal{X}|$. Entropy magically lower bounds the number of bits needed to describe X. Today we'll see that there always exists a function f^* such that $E\ell \circ f^*(X) \leq H(X)$.
2. X such that $X = 0$ w.p. p, $X = 1$ w.p. $1-p$. Last time we saw that a typical sequence of flips should have about np heads (0) and $n(1-p)$ tails (1). What is the probability of observing a "typical" sequence? We saw it was $p_{X^n}(x^n) \approx 2^{-nH(X)}$. This motivated the AEP, a consequence of the weak law of large numbers:

$$-\frac{1}{n}\log p_{X^n}(x^n) \to H(X) \quad \text{in probability.}$$

   This is for any discrete X, not just coin flips. But what does "typical" mean?

The typical set

Definition: The typical set $A_\epsilon^{(n)} \subseteq \mathcal{X}^n$ for p_X is the set of $x^n \in X^n$ for which

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

Properties of $A_\epsilon^{(n)}$:

1. $x^n \in A_\epsilon^{(n)} \Leftrightarrow H(X) - \epsilon \leq -\frac{1}{n}\log p_{X^n}(x^n) \leq H(X) + \epsilon$, by definition.
2. $P(A_\epsilon^{(n)}) = P(X^n \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$, by the AEP.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$. Proof:

$$1 \geq \sum_{x^n \in A_\epsilon^{(n)}} p_{X^n}(x^n) \geq |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)}$$

   since probabilities are bounded by 1.
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ if n is sufficiently large. Proof:

$$(1-\epsilon) \leq P(A_\epsilon^{(n)}) = \sum_{x^n \in A_\epsilon^{(n)}} p_{X^n}(x^n) \leq |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}$$

   where the first inequality holds when n is sufficiently large.

We can imagine the set $A_\epsilon^{(n)}$ sitting inside $\mathcal{X}^n$. The set has $P(A_\epsilon^{(n)}) > 1-\epsilon$ and we have $|A_\epsilon^{(n)}| \approx 2^{nH(X)}$. This is an example of a concentration phenomenon, since basically all of the probability is concentrated on a small subset.

Let's define $B_\delta^{(n)}$ to be the smallest set $\subseteq \mathcal{X}^n$ with probability $\geq 1-\delta$. We can explore how this set relates to $A_\epsilon^{(n)}$, and we'll see that they are almost the same. This underlies the notion that entropy can be considered a measure of "volume."

Claim: $|B_\delta^{(n)}| \approx |A_\epsilon^{(n)}|$. Validation: since $A_\epsilon^{(n)}$ contains most of the probability, any other set with high probability must have a large intersection with $A_\epsilon^{(n)}$.

Proof:

$$1 - \epsilon - \delta \leq P(B_\delta^{(n)} \cap A_\epsilon^{(n)}) = \sum_{x^n \in B_\delta^{(n)} \cap A_\epsilon^{(n)}} p_{X^n}(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H(X)-\epsilon)} \leq |B_\delta^{(n)}| 2^{-n(H(X)-\epsilon)}$$

for large n, where we used the definition of $A_\epsilon^{(n)}$ and then expanded them sum to be over $B_\delta^{(n)}$ instead of the intersection. Now rearrange to get

$$|A_\epsilon^{(n)}| \geq |B_\delta^{(n)}| \geq (1 - \epsilon - \delta) 2^{n(H(X)+\epsilon)} 2^{-n2\epsilon} \geq |A_\epsilon^{(n)}|(1 - \epsilon - \delta) 2^{-n2\epsilon}$$

Now $|B_\delta^{(n)}|$ is sandwiched between two things getting very close since ϵ, δ are small. Thus the sizes of the sets become the same as n gets large.

How to describe X with $H(X)$ bits on average

This is the "typical set encoding," which is not practically useful but is good enough to prove a result.

Step 1: Label all sequences in $A_\epsilon^{(n)}$ from $1, \dots, |A_\epsilon^{(n)}|$. This uses $\log |A_\epsilon^{(n)}| + 1$ bits per label.

Encoding: if $x^n \in A_\epsilon^{(n)}$, add a 1 to the beginning of the typical set label: $f(x^n) = (1, \text{label}(x^n))$. This requires $\log |A_\epsilon^{(n)}| + 2$ bits. If $x^n \notin A_\epsilon^{(n)}$, then label it with a zero and the binary representation of x^n: $f(x^n) = (0, \text{binary}(x^n))$. This requires $n(\log |\mathcal{X}| + 1) + 1$ bits (label each element of $\mathcal{X}$ with a string, we have n of these put together).

This is super sloppy for the atypical sequences, but we don't need to be very careful because those sequences don't have much probability. Let's look at the expected length:

$$\begin{aligned}
\frac{1}{n} E\ell \circ f(X^n) &\leq \frac{1}{n} P(X^n \in A_\epsilon^{(n)})(\log |A_\epsilon^{(n)}| + 2) + \frac{1}{n} P(X^n \notin A_\epsilon^{(n)})(n(\log |\mathcal{X}| + 1) + 1) \\
&\leq \frac{n(H(X) + \epsilon)}{n} + \frac{2}{n} + \delta(n)((\log |\mathcal{X}| + 1) + \frac{1}{n}) \\
&\leq H(X) + 2\epsilon
\end{aligned}$$

This is the first sort of "information limit" that we have derived.

Note that we assume X^n is i.i.d. sampled from the distribution of X. But this is sort of a bad assumption in some cases. Note that our lower bound is absolute:

$$E\ell \circ f(X) \geq H(X) - \log \log 2|\mathcal{X}|$$

for a single sample X. It's "one shot." Later we will see that there is some function such that

$$E\ell \circ f(X) \leq H(X) + 1,$$

even doing away with the i.i.d. assumption.

Definition: Entropy Rate. For a random process $X_1, X_2, \dots = \{X_i\}$, we define

$$H(\{X_i\}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \dots, X_n),$$

provided this limit exists. We'll see that entropy plays the same role for a stochastic process as entropy plays for an i.i.d. process in terms of length of encodings, etc. More next time.

September 8

Next time in Cory 293. The table at the end of HW2 was wrong initially.

Entropy Rate

Definition: Entropy Rate. For a random process $X_1, X_2, \ldots = \{X_i\}$, we define

$$H(\{X_i\}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

provided this limit exists. Note that this reduces to entropy if the X_i are i.i.d.

Shannon-McMillan-Breiman theorem: For stationary ergodic processes,

$$-\frac{1}{n} \log p_{X^n}(X_1, \ldots, X_n) \to H(\{X_i\})$$

with probability 1. Proved in book.

Think of this as generalized AEP for random processes. So, the AEP-like properties carry over from i.i.d. processes.

Stationary means joint probabilities are invariant to shifts in time, so

$$P(X_1, \ldots, X_k) = x_1, \ldots, x_k) = P(X_{l+1}, \ldots, X_{l+k} = x_1, \ldots, x_k).$$

Ergodic means that time averages equal ensemble averages, so

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to EX_i$$

in some sense. So "ergodic" means a law of large numbers type of property should hold.

Lemma: For a stationary process,

$$H(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1}).$$

Proof: First, we need to show the limit exists.

$$0 \le H(X_{n+1}|X_1, \ldots, X_n) \le H(X_{n+1}|X_2, \ldots X_n) = H(X_n \mid X_1, \ldots, X_{n-1})$$

by stationarity. Now, we claim that if $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$ also. We can apply this to what we just saw:

$$a_n = H(X_n \mid X_1, \ldots, X_{n-1}), \quad b_n = \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1}) = \frac{1}{n} H(X_1, \ldots, X_n).$$

Example: Markov chains

Consider the Markov chain with transition matrix

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/3 & 2/3 \\ 1/4 & 1/4 & 2 \end{pmatrix}.$$

The stationary distribution is the distribution π with $\pi P = \pi$. If $X_1 \sim \pi$, then this is a stationary process. To compute the entropy rate, we just need to compute $H(X_n \mid X_{n-1})$ since this process is Markov.

$$H(X_n \mid X_{n-1}) = \sum \pi_i H(X_n \mid X_{n-1} = i) = \sum \pi_i H(P_i)$$

where P_i is the i-th row of P.

Data compression

We've seen that it takes $H(X)$ bits on average to describe X. This is the fundamental idea behind data compression.

Recall from the AEP that it suffices to use roughly $nH(X)$ bits to represent $X_1, \ldots, X_n$ (i.i.d. or stationary and ergodic).

Source code $c : \mathcal{X} \to \{0,1\}^*$.

Figure of merit: $L(c) = \sum p_X(x) \ell \circ c(x)$.

Extension code c^* encodes strings of x's in the obvious way

$$c^*(x_1, \ldots, x_n) = c(x_1)c(x_2)\cdots c(x_n)$$

Nonsingular code (or one-to-one) c satisfies

$$x \neq x' \Rightarrow c(x) \neq c(x')$$

A code c is uniquely decodable if c^* is nonsingular.

A code is prefix-free (or instantaneous) if no codeword is a prefix of any other. Can also say c is a "prefix code."

all codes $\supseteq$ nonsingular codes $\supseteq$ uniquely decodable codes $\supseteq$ prefix-free codes

Kraft inequality:

1. Any prefix code satisfies $\sum_{i=1}^N 2^{-\ell_i} \leq 1$ for $\ell_1, \ldots, \ell_N =$ codeword lengths.
2. If $\sum 2^{-\ell_i} \leq 1$, then $\exists$ prefix code with those codeword lengths.

The second part can be shown with a binary tree using the dyadic probabilities.

First part: put random coin flips into a decoder, wait until some X comes out (if it does). Bound the sum of probabilities of each X_i (which are $2^{-\ell_i}$) by 1, since the X_i are disjoint events here as it's a prefix code.

McMillan Inequality: (to be proved in HW2)
Any uniquely decodable code has codeword lengths satisfying $\sum_{i=1}^n 2^{-\ell_i} \leq 1$.

September 13

Last time, we started talking about data compression using codes $c : \mathcal{X} \to \{0,1\}^*$. We discussed the hierarchy of

all codes $\supseteq$ nonsingular codes (instantaneous codes) $\supseteq$ uniquely decodable codes $\supseteq$ prefix codes

where prefix codes and uniquely decodable codes are most interesting in practice, while instantaneous codes are also interesting in theory.

Last time we proved Kraft's inequality and mentioned the

McMillan Inequality: Any uniquely decodable code satisfies $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$.

Recall: For a one-to-one function f

$$E\ell \circ f(X) = H(X) + D(p_X || q) - \log\left(\sum_x 2^{-\ell \circ f(x)}\right)$$

for $q(x) \propto 2^{-\ell \circ f(x)}$.

Corollary: If $c : \mathcal{X} \to \{0,1\}^*$ is uniquely decodable then

$$E\ell \circ c(X) \geq H(X)$$

How to design good codes? Huffman coding

We want to find

$$\min_{\ell_1,\dots,\ell_N \geq 0} \sum p_X(i) l_i, \quad \text{s.t. } \sum_x 2^{-\ell_i} \leq 1$$

The optimal ℓ_i satisfy

$$\ell_i = \log \frac{1}{p_X(i)}$$

since this makes the inequality tight.

We can't have fractional ℓ_i, so we can just take $\ell_i = \lceil \log \frac{1}{p_X(i)} \rceil$. The inequality we need still holds since we just increased all the ℓ_i. Then 000

$$\sum p_X(i)\ell_i \leq \sum p_X(i) \left(\log \frac{1}{p_X(i)} + 1\right) = H(X) + 1$$

Thus our Kraft proof tree construction gives a code that's within 1 bit of the entropy! This seems pretty good. This gives us:]

Theorem: Any uniquely decodable code satisfies

$$E\ell \circ c(X) \geq H(X)$$

Moreover, $\exists c^*$ such that $E\ell \circ c^*(X) \leq H(X) + 1$.

This is the first example of an "information-theoretic result" that we've seen. It includes a converse (an impossibility) and an achievability. We call this an information limit (since it has an upper and lower bound).

Huffman codes. An "efficient" algorithm for constructing the best prefix code. Best described by pictures (build up a tree).

Construction:

1. Pick two smallest probabilities and merge. Label the branches 0 and 1.
2. Repeat until all probabilities merged together.
3. Read off codewords by tracing back in the tree.

Theorem: Huffman coding is optimal among all uniquely decodable codes. (Proved in book, not in class.) That is, if c^* is a Huffman code, then

$$E\ell \circ c^*(X) \leq E\ell \circ c(X)$$

for any other uniquely decodable code c.

Shannon-Fano-Elias coding and arithmetic coding

But... can we get rid of the annoying "+1"? (comment about wedding invite) The answer is to group independent symbols together and code $c : \mathcal{X}^n \to \{0,1\}^*$. Then

$$H(X^n) \leq E\ell \circ^* (X^n) \leq H(X^n) + 1 = nH(X^n) + 1$$

so our 1 becomes $\frac{1}{n}$ when measuring bits per symbol. But if we take n to be large, we have to build a tree with $|\mathcal{X}|^n$ leaves.

This motivates arithmetic coding, which has complexity linear in n. To get to this, we first need to describe *Shannon-Fano-Elias coding*. We will see that

$$E\ell \circ c_{SFE}(X) \leq H(X) + 2$$

but this flexibility allows us to scale more gracefully with n.

Suppose we have $p_X = (p_1, \ldots, p_n)$. We partition the interval $[0, 1)$ into n intervals of the form

$$\left[\sum_{j=1}^{i-1} p_j, \sum_{j=1}^{i} p_j\right)$$

for $i = 1, \ldots, n$. To encode symbol i, look at the binary representation of the midpoint of interval i and truncate it to $\lceil \log \frac{1}{p_i} \rceil + 1$ bits. This number is guaranteed to lie in interval i, because truncation subtracts at most

$$2^{-(\lceil \log \frac{1}{p_i} \rceil + 1)} \leq 2^{-\left(\log \frac{1}{p_i} + 1\right)} = \frac{p_i}{2}$$

from the midpoint.

I stopped paying attention for arithmetic coding. Oops! You just write out the intervals, then recursively zoom into the interval for each symbol in order. Then you truncate the message in binary to

$$\left\lceil \log \frac{1}{p_{X^n}(X^N)} \right\rceil + 1$$

bits.

September 15

“Today, you guys are in for a real treat.”

The Communication Problem

“The cornerstone result of information theory.” What is the fundamental process of communication?

msg $\to$ transmitter $\to$ signal $\sim$ channel (+noise corruption!) $\sim$ signal $\to$ receiver $\to$ msg

“Kind of like ESP” (transmitting thoughts from your brain to my brain using speech understanding).

Assumptions:

A1) Messages are random variables uniformly distributed over $\{1, \ldots, M\}$.

Justification: bits should look like fair coin flips. If there’s any bias or correlation, then the message could be compressed further.

A2) The channel has known* statistical properties, so that if X goes in and Y comes out, we know the conditional distribution $P_{Y|X}$.

Justification: In real life, we can take measurements.

These are mild assumptions and we have very solid justifications.

In general, we can describe the channel by $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$: an input and output alphabet and a conditional distribution.

Discrete-time, discrete-alphabet, memoryless channel

$(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ such that $\mathcal{X}, \mathcal{Y}$ are discrete (finite) sets and we send one element of $\mathcal{X}$ through at a time, with the statistics staying constant over time.

Example: Binary Symmetric Channel with crossover probability p (i.e., BSC(p)).

You input 0 or 1, and it outputs the same bit with probability $1 - p$ and the opposite bit with probability p.

“I whisper bits to Carlos, and he lies 10% of the time.”

Here

$$p_{Y|X}(y|x) = 1\{y \neq x\}p + 1\{y = x\}(1 - p)$$

Now, let’s talk about the transmitter. It takes the message and converts it to something suitable for going over a channel.

Definition: an (M, n)-block code for the channel $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ uses n symbols from $\mathcal{X}$ to represent (or really, transmit) one of M messages $W_i, i = 1, \ldots, M$.

So W gets converted into $(X_1, \ldots, X_n) = X^n(W)$ by the code, then goes through channel $P_{Y^n|X^n}$ to get $(Y_1, \ldots, Y_n) = Y^n$ then gets decoded to get $\hat{W}(Y^n)$.

Definition: the rate of an (M, n) code is

$$R = \frac{\log M}{n} = \frac{\text{\# of information bits}}{\text{\# of channel uses}}$$

Note that an (M, n) code is a $(2^{nR}, n)$ code.

A rate R is *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that

$$\max_{i \in 1,\ldots,2^{nR}} P(\hat{W} \neq W_i \mid W = W_i) \to 0 \quad \text{as } n \to \infty,$$

which we call *reliable communication.*

It's not clear that we can send information at a constant positive rate and vanishing probability of error? It's not an easy question to answer immediately.

Definition: "Operational" definition of channel capacity.

$$C(P_{Y|X}) = C = \sup\{R : R \text{ is achievable}\}.$$

" These are some of the most important definitions in the whole class."

By definition:

1) All rates $R < C$ are achievable, i.e., for any $\epsilon > 0, \exists n$ s..t there is a $(2^{nR}, n)$ code with $P(\hat{W} \neq W_i \mid W = W_i) < \epsilon$, $\forall i = 1, \ldots, 2^{nR}$.
2) No rates $R > C$ are achievable, i.e., $\exists c > 0$ such that for any sequence of $(2^{nR}, n)$ codes will have

$$\liminf_{n \to \infty} P(\hat{W} \neq W_i \mid W = W_i) > c > 0$$

Theorem: Shannon's Channel Coding Theorem

$$C = \max_{p_X} I(X; Y)$$

Properties of C

1. $C \geq 0$. Otherwise it would be weird.
2. $C \leq \log |\mathcal{X}|$ and $C \leq \log |\mathcal{Y}|$. Why?

$$I(X; Y) = H(X) - H(X \mid Y) \leq H(X) \leq \log |\mathcal{X}|$$

Do it for Y to get the other bound. This is possible in the ideal case, when there is no noise.

3. Recall that $I(X;Y)$ is concave in p_X for fixed $p_{Y|X}$ (which is fixed for a given channel). Computing C is a convex optimization problem!

Example 1: BSC(p). We can compute this by hand.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(p) \leq 1 - H(p)$$

We can achieve this upper bound using $X \sim \text{Bern}(1/2)$, which gives Y the same distribution and thus $H(Y) = 1$. So $C = 1 - H(p)$. "Beautiful! Not even ugly."

Example 2: BEC(p). Put in 0 or 1, what comes out is 0, 1, or e (erased!). e has probability p, and transmitting correctly has probability $1 - p$.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(p)$$

Let E be a variable that is 1 if $Y = e$ and 0 if $Y \neq e$.

$$H(Y) = H(Y,E) = H(E) + H(Y|E) = H(p) + (pH(Y|E=1) + (1-p)H(Y|E=0))$$

We have $H(Y|E = 1) = 0$ and $H(Y|E = 0) \leq 1$. Turns out that equality holds and we have $C = 1 - p$.

September 20

The communication problem: generally, we have some message $W \in \{1, \dots, 2^{nR}\}$. An encoder maps to $X^n(W)$. A channel $\prod P_{Y_i|X_i}$ maps to Y^n. A decoder estimates $\hat{W}(Y^n)$.

Last time, we talked about achievable rates. A rate R is achievable if $\exists$ a sequence of $(2^{nR}, n)$ codes such that

$$\max_i P(\hat{W} \neq W_i \mid W = W_i) \to 0 \text{ as } n \to \infty.$$

The channel capacity C is the supremum of achievable rates, i.e., the maximum rate at which information can be transmitted reliably. We saw the theorem (not proved yet) that $C = \max_{p_X} I(X;Y)$.

A heuristic explanation for $BEC(p)$

Let's first argue an upper bound on achievable rates R. One way to upper bound is to make the transmitter more capable. From the point of view of the transmitter, you know the decoder will see something like $e\_\_e\_\_\_ee\_$. If there are n total symbols, then we will have roughly pn erasures. If we are omniscient and know when erasures will occur, we can send at most $n(1-p)$ bits with n transmissions. This tells us that any achievable R is at most $1 - p$.

How would you ever send information at this rate? Consider a binary matrix G with dimensions $n \times (1-p-\epsilon)n$. Let's assume G has full rank (over addition mod 2). Let $X^n(W) = GW$, where W is $(1-p-\epsilon)n$ bits long. The channel will produce roughly pn erasures, leaving a vector Y^n with roughly $(1-p)n$ elements not erased. Now we have a system of equations to solve: $GW = Y$ with $(1-p)n$ and $(1-p-\epsilon)n$ unknowns!

One more example of computing channel capacity: "noisy typewriter"

The noisy typewriter channel on k letters maps letter i to i w.p. $1/2$ and $i+1 \mod k$ w.p. $1/2$.

Our strategy is to upper bound $I(X;Y)$, then cleverly choose a distribution that achieves that upper bound.

$$I(X;Y) = H(Y) - H(Y \mid X) \leq \log k - 1.$$

This is achieved by the uniform distribution on the $k/2$ letters $\{0, 2, 4, \ldots, k-2\}$. Thus $C = \log k - 1$.

A preview of achievability in the channel coding theorem: All channels look like noisy typewriters for sufficiently large n.

Proving the converse of channel capacity

Proof of "converse:" If R is achievable, then $R \leq C$.

If R is achievable, there exists a $(2^{nR}, n)$ code with $\max_i P(\hat{W}^{(n)} \neq W_i \mid W^{(n)} = W_i) = \epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. This implies that $P(\hat{W}^{(n)} \neq W^{(n)}) \leq \epsilon_n$.

$$\begin{aligned}
nR &= H(W^{(n)}) \\
&= H(W^{(n)} \mid \hat{W}^{(n)}) + I(W^{(n)}; \hat{W}^{(n)}) \\
&\leq 1 + \epsilon_n \log 2^{nR} + I(W^{(n)}; \hat{W}^{(n)}) \quad \text{by Fano's ineq.} \\
&= 1 + \epsilon_n nR + I(W^{(n)}; \hat{W}^{(n)}) \\
&\leq 1 + \epsilon_n nR + I(X^n(W^{(n)}); Y^n) \quad \text{by data processing ineq.} \\
&= 1 + \epsilon_n nR + H(Y^n) - H(Y^n \mid X^n(W)) \\
&\leq 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - H(Y^n \mid X^n(W)) \\
&= 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X^n, Y_{i-1}) \\
&= 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X_i) \quad \text{by DMC property} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^n I(X_i; Y_i) \\
&\leq 1 + \epsilon_n nR + nC.
\end{aligned}$$

We conclude that

$$nR \leq 1 + \epsilon_n nR + nC \quad \Rightarrow \quad R \leq \frac{1}{n} + \epsilon_n R + C \to C.$$

Note that this result still holds even if the probability of error is only bounded *on average*.

We can rearrange our inequality to get

$$\epsilon_n \geq 1 - \frac{C}{R} - \frac{1}{nR} \to 1 - \frac{C}{R}$$

which strictly bounds the probability of error to be positive if $R > C$. This is a weak converse. There's a strong converse that says the probability of error goes to 1 exponentially fast when $R > C$.

September 22

Joint Typical Sequences

The set $A_\epsilon^{(n)}$ (or $(A_\epsilon^{(n)}(X,Y))$ of jointly typical sequences $X^n, Y^{,}$ with respect to a distribution P_{XY} is the set of n-sequences with empirical entropies close to true entropy:

$$\begin{gathered} A_\epsilon^{(n)} = \left\{ (x^n, y^n) : \left| -\frac{1}{n} \log P_{X^n,Y^n}(x^n, y^n) - H(X,Y) \right| < \epsilon \right\} \bigcap \\ \left\{ (x^n, y^n) : \left| -\frac{1}{n} \log P_{X^n}(x^n) - H(X) \right| < \epsilon \right\} \bigcap \\ \left\{ (x^n, y^n) : \left| -\frac{1}{n} \log P_{Y^n}(Y^n) - H(Y) \right| < \epsilon \right\} \end{gathered}$$

That is, all elements must appear both jointly and marginally typical.

Joint AEP: Let $X^n, Y^n \sim \prod P_{XY}(x_i, y_i)$.

1. $P((X^n, Y^n) \in A_\epsilon^{(n)}(X,Y)) \to 1$.
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. If $(\tilde{X}^n, \tilde{Y}^n) \sim P_{X^n} P_{Y^n}$, then

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

   So if the points are sampled from the marginals independently, then the probability of being in the typical set is vanishingly small.

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

   So not all pairings of typical X^n sequences and typical Y^n sequences gives a typical (X^n, Y^n) sequence. In fact, most do not.

Achievability

Claim: All rates $R < C$ are achievable. The proof is non-constructive, uses the probabilistic method.

Our set is $\mathcal{C} = (2^{nR}, n)$-codes. The desired property is $P =$ code has small probability of error.

Random Coding: Fix $P_X, \epsilon > 0$. Generate a $(2^{nR}, n)$-code at random according to P_X (so $X_i(w) \sim$ i.i.d. from P_x):

$$\mathbb{C} = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ X_1(2) & X_2(2) & \cdots & X_n(2) \\ \vdots & & & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{bmatrix}$$

The encoding scheme: for message W, send $X^n(w) = w^{th}$ row of $\mathbb{C}$.

So, $\mathbb{P}(\mathbb{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} P_X(X_i(W))$.

Typical Set Decoding: Receiver declares $\hat{w}$ was sent if

1. $(X^n(\hat{w}), Y^n) \in A_\epsilon^{(n)}(X, Y)$.
2. There is no other index $k \neq \hat{w}$ with $(X^n(k), Y^n) \in A_\epsilon^{(n)}(X, Y)$.

Otherwise, the receiver declares an error.

Strategy: compute expected probability of error for this scheme, averaging over random codebooks $\mathbb{C}$ drawn according to the distribution $\mathbb{P}(\mathbb{C})$.

The probability of error over the randomness in the channel (for a fixed code) is

$$\lambda_i(\mathbb{C}) = P(\hat{W} \neq i \mid X^n = X^n(i))$$

The average probability of error for a code $\mathbb{C}$:

$$P_e^{(n)}(\mathbb{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathbb{C})$$

The average for all codes is

$$\begin{aligned} P_{error} &= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) P_e^{(n)}(\mathbb{C}) \\ &= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathbb{C}) \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \lambda_w(\mathbb{C}) \\ &= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \lambda_1(\mathbb{C}) \quad \text{by symmetry} \\ &= P(\mathcal{E} \mid W = 1) \end{aligned}$$

i.e., the probability of error averaged over all codes, conditioned on message 1 being sent.

Define error events $E_i = \{(X^n(i), Y^n) \in A_\epsilon^{(n)}(X, Y)\}$ for $i = 1, \ldots, 2^{nR}$.

Then

$$P(\mathcal{E} \mid W = 1) = P(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}} \mid W = 1)$$

since either the pair isn't in the typical set or it is in the wrong typical set.

$$\begin{aligned}
P(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}} \mid W = 1) &\leq P(E_1^c \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1) \\
&= P((X^n, Y^n) \notin A_\epsilon^{(n)}) + \sum_{i=2}^{2^{nR}} P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \\
&\leq \epsilon + 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} \\
&\leq 2\epsilon
\end{aligned}$$

for sufficiently large n, if $R < I(X;Y) - 3\epsilon$.

We have just shown: For any fixed $R < C, \epsilon > 0$, there exists a $(2^{nR}, n)$-code with average probability of error $< \epsilon$.

September 29

First midterm is next Tuesday (October 4). Can bring one page (one side) of notes. Topics are: information measures, entropy rate, source coding, channel capacity, and the channel coding theorem.

Wrapping up channel coding

Our theorem was $C = \max_{p_X} I(X;Y)$.

Converse: IfR achievable, $R < C$ with maximal probability of error $\max_i P(\hat{W} \neq W_i \mid W = W_i)$ or average probability of error $P(\hat{W} \neq W)$.

Achievability: There exist $(2^{nR}, n)$-codes with $P(\hat{W} \neq W) \to 0$ as $n \to \infty$, provided $R < C$. But this is a bound on average error. However, we can fix this simply using Markov's inequality.

We have

$$P(\hat{W} \neq W) = \sum_{i=1}^{2^{nR}} P(\hat{W} \neq W_i \mid W = W_i) P(W = W_i) = E[P(\hat{W} \neq W_i \mid W)]$$

$$\#\{i : P(\hat{W} \neq W_i \mid W = W_i) > \lambda\} \leq \frac{2^{nR} P(\hat{W} \neq W)}{\lambda}$$

Take $\lambda = 2P(\hat{W} \neq W)$. Then

$$\#\{i : P(\hat{W} \neq W_i \mid W = W_i) > 2P(\hat{W} \neq W)\} \leq \frac{1}{2} 2^{nR}$$

Throw these messages away. The new rate is

$$\frac{\log(\#\text{ codewords})}{n} = \frac{\log 2^{nR-1}}{n} = R - \frac{1}{n}$$

So the new codes have maximum probability of error $2P(\hat{W} \neq W) \to 0$ and rate $R-1/n \to R$.

Two questions

Suppose you have a channel but you get feedback about what Y value was send by $P_{Y|X}$.

Question: what is capacity with feedback?

Suppose you have X^n and Y^n (dependent on X^n) two codes sending nR_X and nR_Y bits versus a single code sending nR bits where both X^n, Y^n are known.

Question: how much worse is the cooperating code (can it beat rate $H(X)+H(Y)$ versus the lower bounding rate $H(X,Y)$ of the single code)? We'll do this one. Done in 1973, is in chapter 15.4 of the book.

Lossless coding of correlated sources

R_X, R_Y are achievable if there exists $f : \mathcal{X}^n \to \{1, \dots, 2^{nR_X}\}, g : \mathcal{Y}^n \to \{1, \dots, 2^{nR_Y}\}$ and $\phi : \{1, \dots, 2^{nR_X}\} \times \{1, \dots, 2^{nR_Y}\} \to \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$P(\phi(f(X^n), g(Y^n) \neq X^n, Y^n) \to 0$$

Achievable rate region: the closure of achievable rates (where closure is the analog of supremum for multiple dimensions).

Feedback channel

Never mind, we'll switch to this problem.

I have a channel with feedback, so my input X_i at time instant i can depend on $W, Y_1, \dots, Y_{i-1}, X_1, \dots, X_{i-1}$.

Theorem: $C_{FB} = C$.

Feedback doesn't help. However, it can simplify things (e.g., in the BEC example). It also can get us to capacity much more quickly. Without feedback, the probability of error of the best code is approximately 2^{-nE}, but with feedback, sometimes we can get $2^{-2^{nE}}$.

Proof: In the converse, we had

$$\begin{aligned} nR = H(W) &\leq I(W; \hat{W}) + n\epsilon_n \quad \text{Fano} \\ &\leq I(W; Y^n) + n\epsilon_n \quad \text{DPI} \end{aligned}$$

but the later step

$$H(Y^n \mid X^n) = \sum H(Y_i \mid X_i)$$

fails since we have added feedback ($Y_i \mid X_i$ are not conditionally independent variables any more). This time,

$$\begin{aligned}
I(W;Y^n) + n\epsilon_n &= H(Y^n) - H(Y^n \mid W) + n\epsilon_n \\
&\leq \sum (H(Y_i) - H(Y_i \mid W, Y^{i-1})) + n\epsilon_n \\
&= \sum (H(Y_i) - H(Y_i \mid W, Y^{i-1}, X_i)) + n\epsilon_n \qquad \text{since } X_i = f(W, Y^{i-1}) \\
&= \sum (H(Y_i) - H(Y_i \mid X_i)) + n\epsilon_n.
\end{aligned}$$

Now the rest of the proof matches up.

October 6

"I like the average to be about 50% because then the... entropy is high, which reveals lots of... information."

Midterm 1 postmortem: problem 4

See online solutions.

Continuous random variables

Up until now, we've always had $\mathcal{X}$ discrete. In this case

$$H(X) = -\sum p_X(x) \log p_X(x)$$

Question: What if X is a continuous random variable? We have "differential entropy:"

$$h(X) - \int f_X(x) \log f_X(x) dx = E \log \frac{1}{f_X(X)}$$

where f_X is the density of X.

Example: $X \sim U[0, a]$. Here $h(X) = \log a$. Note that this is no longer bounded below by 0. "Shattered my whole world."

Why is this OK? Let's look at the Riemann sum:

$$h(X) = -\int f(x) \log f(x) dx \approx -\sum_{i=-\infty}^{\infty} \frac{1}{N} f(i/N) \log f(i/N)$$

We can think of this as a discrete random variable $[X]_N$ that can fall in N buckets. The mass function will be about $\frac{1}{N} f(i/N)$ for the i-th bucket.

$$= -\sum_{i=-\infty}^{\infty} \frac{1}{N} f(i/N) \log \frac{1}{N} f(i/N) - \log N = H([X]_N) - H([U]_N)$$

So it's the differential between my distribution and the uniform distribution on $[0,1]$, both discretized to the same level. The entropy of the density can surpass that of the uniform distribution because it is defined over the whole line, not just $[0,1]$.

"None of these questions are particularly relevant."

Example:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

"This is another one of those miracles of life."

$$\begin{aligned} h(X) &= \int f(x) \log \frac{1}{f(x)} dx \\ &= \frac{1}{2} \log 2\pi\sigma^2 + \int f(x) \frac{x^2}{2\sigma^2} \log e dx \\ &= \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \log e \\ &= \frac{1}{2} \log 2\pi e\sigma^2 \end{aligned}$$

Joint Differential Entropy

$$h(X_1, \ldots, X_n) = -\int f \log f$$

for $f =$ joint density of $X_1, \ldots, X_n$.

Example:

$$f(x) = \frac{1}{(2\pi)^{n/2}|K|^{1/2}} \exp\left(-\frac{(x-\mu)^T K^{-1}(x-\mu)}{2}\right)$$

This is $N(\mu, K)$.

$$\int f \log \frac{1}{f} = \frac{1}{2} \log(2\pi)^n |K| + \frac{1}{2} \log(e) E(x-\mu)^T K^{-1}(x-\mu)$$

Use trace trick to see

$$\operatorname{tr}(x-\mu)^T K^{-1}(x-\mu) = \operatorname{tr}(x-\mu)(x-\mu)^T K^{-1}$$

which is n in expectation, so we get

$$h(X) = \frac{1}{2} \log(2\pi e)^n |K|.$$

Conditional DE

$$h(Y \mid X) = \int f_{XY}(x,y) \log \frac{1}{f_{Y|X}(y|x)} dy dx = -E \log \frac{1}{f_{Y|X}(Y|X)}$$

The chain rule continues to hold, so $h(Y, X) = h(X) + h(Y \mid X)$.

Relative Entropy

$$D(f||g) = \int f \log \frac{f}{g} \geq 0$$

as long as $\mathrm{supp}(f) \subseteq \mathrm{supp}(g)$. It's still convex.

Mutual information

$$I(X;Y) = D(f_{XY}||f_X f_Y) = h(X) - h(X \mid Y) \geq 0$$

Thus $h(X) \geq h(X \mid Y)$ is still true.

The chain rule still holds.

$h(X+c) = h(X)$ for c any constant.

$h(cX) = h(X) + \log|c|$, since the density scales by $\frac{1}{|c|}$ when we go from X to cX.

For vector valued X and a matrix A, $h(AX) = h(X) + \log|A|$.

"Want to know one of the most beautiful results in all of information theory? It's not even in my notes right now, I'll just tell it to you."

Entropy Power Inequality: X, Y independent random vectors on $\mathbb{R}^n$.

$$2^{\frac{2}{n}h(X+Y)} \geq 2^{\frac{2}{n}h(X)} + 2^{\frac{2}{n}h(Y)}$$

October 13

Entropy power inequality

(in one dimension, adapted from Rioul on arXiv in July)

$$2^{2h(U+V)} \geq 2^{2h(U)} + 2^{2h(V)}$$

Let $U = \sqrt{\lambda}X, V = \sqrt{1-\lambda}Y$ so that this becomes equivalent to

$$2^{2h(\sqrt{\lambda}X+\sqrt{1-\lambda}Y)} \geq \lambda 2^{2h(X)} + (1-\lambda)2^{2h(Y)}$$

since $h(\sqrt{\lambda}X) = h(X) + \frac{1}{2}\log\lambda$. Using Jensen on the right side and taking logs, we get Lieb's inequality:

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h(X) + (1-\lambda)h(Y)$$

If you plug in $\lambda = 2^{2h(X)}/(2^{2h(X)} + 2^{2h(Y)})$, you get exactly the entropy power inequality.

Let's assume the densities of X, Y are non-vanishing. (This is OK since if it's not true, we can convolve with a skinny Gaussian and take the limit.) Let Φ be the CDF of $N(0,1)$, F_X be the CDF of X, F_Y be the CDF of Y.

Let $T_X(x^*) = F_X^{-1}(\Phi(x^*))$. If $X^* \sim N(0,1)$, then $T_X(X^*)$ has the distribution of X. This function is increasing since it is the composition of increasing functions. Define T_Y similarly.

Let X^*, Y^* be i.i.d. $N(0,1)$. Then $T_X(X^*), T_Y(Y^*) = X, Y$ in distribution. Also take $\tilde{X}, \tilde{Y} \sim$ i.i.d. $N(0,1)$. We write

$$X^* = \sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}, \quad Y^* = \sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}Y$$

since the Gaussian is rotationally invariant.

Take

$$\Theta_{\tilde{Y}}(\tilde{X}) = \sqrt{\lambda} T_X(X^*) + \sqrt{1-\lambda} T_Y(Y^*)$$

Then

$$\frac{d}{d\tilde{x}} \Theta_{\tilde{y}}(\tilde{x}) = \lambda T_X'(x^*) + (1-\lambda) T_Y'(y^*)$$

Now if $f =$ density of $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$, by the change of variables formula,

$$f_{\tilde{y}}(\tilde{x}) = f(\Theta_{\tilde{y}}(\tilde{x}))\Theta_{\tilde{y}}'(\tilde{x}) \quad \Rightarrow \quad \lambda T_X'(x^*) + (1-\lambda)T_Y'(y^*) > 0$$

and where $f_{\tilde{y}}(\tilde{x})$ is a density in $\tilde{x}$.

$$\begin{aligned}
h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= E \log \frac{1}{f(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)} \\
&= E \log \frac{1}{f(\Theta_{\tilde{Y}}(\tilde{X}))} \\
&= E \log \frac{\Theta_{\tilde{Y}}'(\tilde{X})}{f_{\tilde{Y}}(\tilde{X})} \\
&= h(\tilde{X}) + E \log \frac{g(\tilde{X})}{f_{\tilde{Y}}(\tilde{X})} + E \log \Theta_{\tilde{Y}}'(\tilde{X}) \quad \text{ for } g = N(0,1) \text{ density} \\
&= h(\tilde{X}) + E_{\tilde{Y}}\left[E_{\tilde{X}} \left[\log \frac{g(\tilde{X})}{f_{\tilde{Y}}(\tilde{X})} \mid \tilde{Y} \right]\right] + E \log \Theta_{\tilde{Y}}'(\tilde{X}) \\
&\geq h(\tilde{X}) + E \log \Theta_{\tilde{Y}}'(\tilde{X}) \quad \text{ since we removed a KL divergence} \\
&= h(g) + E \log \Theta_{\tilde{Y}}'(\tilde{X}) \\
&\geq h(g) + \lambda E \log T_X'(\tilde{X}) + (1-\lambda) E \log T_Y'(\tilde{Y}) \quad \text{ by Jensen's and concavity of log} \\
&= \lambda \left(h(\tilde{X}) + E \log T_X'(\tilde{X}) \right) + (1-\lambda)\left(h(\tilde{Y}) + E \log T_Y'(\tilde{Y}) \right) \\
&= \lambda h(X) + (1-\lambda) h(Y) \quad \text{ by change of variables}
\end{aligned}$$

To explain the last step:

$$h(\tilde{X}) + E \log T_X'(\tilde{X}) = E \log \frac{T_X'(\tilde{X})}{g(\tilde{X})} = E \log \frac{1}{f_X(T_X(\tilde{X}))} = E \log \frac{1}{f_X(X)}$$

since $g(\tilde{X}) = f_X(T_X(\tilde{X}))T_X'(\tilde{X})$.

“It’s amazing, now you can go home and prove the entropy power inequality to your friends.”

Extensions

The inequality is dimension-free: no constants are introduced as n increases.

We just proved
$$2^{2h(X+Y)} \geq 2^{2h(X)} + 2^{2h(Y)}$$

The *conditional EPI* says that if X, Y are conditionally independent given U,
$$2^{2h(X+Y|U)} \geq 2^{2h(X|U)} + 2^{2h(Y|U)}$$

The basic idea behind the proof: by the EPI,
$$2h(X+Y \mid U=u) \geq \log(2^{2h(X|U=u)} + 2^{2h(Y|U=u)})$$

then apply the convexity of log-sum-exp.

For X^n, Y^n independent random variables on $\mathbb{R}^n$,
$$2^{\frac{2}{n}h(X^n+Y^n)} \geq 2^{\frac{2}{n}h(X^n)} + 2^{\frac{2}{n}h(Y^n)}$$

We can prove this with the chain rule, since
$$h(X^n+Y^n) = h(X_n+Y_n) + h(X^{n-1}+Y^{n-1} \mid X_n+Y_n)$$

Then
$$\frac{2}{n}h(X^{n-1}+Y^{n-1} \mid X_n+Y_n) \geq \frac{n-1}{n}\log(2^{\frac{2}{n-1}h(X^{n-1}|X_n)} + 2^{\frac{2}{n-1}h(Y^{n-1}|Y_n)})$$

and
$$\frac{2}{n}h(X_n+Y_n) \geq \frac{1}{n}\log(2^{2h(X_n)} + 2^{2h(Y_n)})$$

Use the convexity of log-sum-exp to add these two and get
$$\frac{2}{n}h(X^n+Y^n) \geq \log(2^{\frac{2}{n}h(X^n)} + 2^{\frac{2}{n}h(Y^n)})$$

Something else

“Ooh, maybe I could do something else.”

“Let’s say we’re playing a game. Do you guys know game theory? Doesn’t matter.”

“I have an information theory t-shirt with that inequality, but it’s like, you’re at the pizza shop, and the pizza guy is like, ‘what does that mean?’ And you’re like ‘...I just wanna eat my pizza, man.’ ”

You have an additive Gaussian channel. Put in $X \sim P_X$, add $Z \sim P_Z$, get $Y = X + Z$.

Say you have an adversary who’s choosing the worst P_Z.
$$\sup_{P_X} \inf_{P_Z} I(X; X+Z)$$

We need a constraint. Take $EX^2 < \sigma_X^2$ and $EZ^2 \leq \sigma_Z^2$.

First, write $I(X; X+Z) = h(X+Z) - h(Z)$.

$$\begin{aligned} h(X+Z) - h(Z) &\geq h(X+Z) - h(Z^*) \\ &\geq h(X^*+Z^*) - h(Z^*) \\ &= \frac{1}{2}\log(1+\frac{\sigma_X^2}{\sigma_Z^2}). \end{aligned}$$

In the other direction,

$$\begin{aligned} \sup_{P_X}\inf_{P_Z} h(X+Z) - h(Z) &\leq \sup_{P_X} h(X+Z^*) - h(Z^*) \\ &= \frac{1}{2}\log(1+\frac{\sigma_X^2}{\sigma_Z^2}). \end{aligned}$$

by Gaussian channel capacity.

Thus

$$\sup_{P_X}\inf_{P_Z} I(X; X+Z) = \inf_{P_Z}\sup_{P_X} I(X; X+Z) = \frac{1}{2}\log(1+\frac{\sigma_X^2}{\sigma_Z^2}).$$

So there's an equilibrium where we are both at Gaussians (called a saddle point property).

Further,

$$I(X; X+Z^*) \leq I(X; X+Z) \leq I(X^*; X^*+Z^*)$$

Gaussian noise is worst case.

October 18

Lossy compression: distortion

We've seen $X^n \to$ encoder $\overset{nR \text{ bits}}{\longrightarrow}$ decoder $\to \hat{X}^n$.

If $R > H(X)$, then $\hat{X}^n = X^n$ is possible (with high probability).

Q: If $R < H(X)$, then what? Is the problem hopeless?

Distortion function (measure): we need to have some measure of "fidelity" to judge how good our recovery is.

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+, \quad d_{max} = \max_{\mathcal{X},\hat{\mathcal{X}}} d(x,\hat{x}) < \infty.$$

We call this bounded distortion. An example of unbounded could be $(x-\hat{x})^2$ (squared error or quadratic loss). Our theory will apply to bounded.

Example: Hamming distortion.

$$d(x,\hat{x}) = 1\{x \neq \hat{x}\}$$

Has the property that $E[d(X,\hat{X})] = P(X \neq \hat{X})$.

Distortion function between sequences: This is just the average per-symbol distortion

$$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i).$$

Rate-distortion theorem

$f_n : \mathcal{X}^n \to [1 : 2^{nR}]$ = encoding function.

$g_n : [1 : 2^{nR}] \to \hat{\mathcal{X}}^n$ = decoding function.

What is this scheme's performance? The figure of merit is the expected distortion,

$$E[d(X^n, g_n(f_n(X^n)))] = \sum_{x^n} p_{X^n}(x^n) d(x^n, g_n(f_n(x^n)))$$

where $g_n(f_n(X^n)) = \hat{X}^n$.

A "rate-distortion" pair (R, D) is achievable if there exists a sequence of $(2^{nR}, n)$-codes (f_n, g_n) such that

$$\lim_{n\to\infty} E[d(X^n, g_n \circ f_n(X^n))] \le D$$

The definition is equivalent to saying that $\forall \epsilon > 0$,

$$P(Ed(X^n, g_n \circ f_n(X^n)) > D + \epsilon) \to 0.$$

We can plot a rate-distortion region. This region extends up and out and is convex (since you can use a convex combination of rate-distortion pairs). We can consider the lower boundary of this region. On the boundary, the point with zero distortion has rate $H(X)$, and the point with zero rate has distortion bound $\min_{\hat{x}} Ed(X, \hat{x})$. We call this boundary (specifically, the closure of the region) the Rate-Distortion Function $R(D)$.

"By the way, Shannon did not solve this problem in his first paper. He solved this problem in his second paper."

Theorem: For $\mathcal{X}^n \sim$ i.i.d. p_X and bounded distortion,

$$R(D) = \min_{\substack{p_{\hat{X}|X}: \\ Ed(X,\hat{X}) \le D}} I(X; \hat{X})$$

We minimize rather than maximizing (as in channel coding) since we want to send as few bits as possible. Also, this case is constrained since we can only select channels that satisfy our distortion requirement. This is a convex optimization problem (mutual information is convex in the conditional distribution for a fixed marginal and concave in the marginal for a fixed conditional), just like channel coding.

Remember: channel coding can be considered a packing problem, since we pack the space $\mathcal{Y}^n$ with as many copies of the range of the channel given $\mathcal{X}^n$ as we can without overlapping too much (then the channel is basically invertible).

In lossy compression, there are only 2^{nR} possible reconstructed sequences $\hat{X}^n$ (because the rate is R). We can consider the balls of X^n such that $d(X^n, \hat{X}^n) \le D$ for each $\hat{X}^n$, so the problem for rate distortion is to select enough $\hat{X}^n$ so that the union of the distortion balls covers the entire space. This is a covering problem.

Example:

$$X = \begin{cases} 1 & \text{w.p. } p \le 1/2 \\ 0 & \text{w.p. } 1 - p \end{cases}$$

We want to lower bound $I(X;\hat{X})$ s.t. $Ed(X,\hat{X}) \leq D$.

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X \mid \hat{X}) \\ &= H(p) - H(X \oplus \hat{X} \mid \hat{X}) \\ &\geq H(p) - H(X \oplus \hat{X}) \\ &\geq H(p) - H(D) \end{aligned}$$

where $\oplus$ is XOR. This turns out to be the mutual information corresponding to a binary symmetric channel with transition probability D and output distribution $(p, 1-p)$.

We can achieve this lower bound with $p(\hat{X}=0) = \frac{1-p-D}{1-2D}$.

Thus $R(D) = H(p) - H(D)$ for $D \leq p$ and 0 for $D > p$.

October 20

Midterm statistics: mean 23.7, standard deviation 10.2, maximum 42 (out of 60, or out of 36 without the final problem).

"It would be a great first problem for exam number two. The first problem of exam number one. Then I would know exactly who looked at the solution."

Rate-distortion function, coding theorem examples

I send $nR \leq H(X)$ bits to encode X^n. The encoder is f_n, and the decoder is g_n. We want

$$Ed(X^n, \hat{X}^n) \leq D$$

so that the estimate is close to the source on average.

Theorem:

$$R(D) = \min_{\substack{p_{\hat{X}|X}: \\ Ed(X,\hat{X}) \leq D}} I(X;\hat{X})$$

Again, it's one of those characterizations where beyond this is impossible but I can get arbitrarily close (like the channel coding theorem).

Example: X∼Bern(p) and d = Hamming distance. $R(D) = (H(p) - H(D))^+$.

Example: We were assuming finite source alphabets and bounded distortion functions.

"What kind of information theorist can't draw a Gaussian..."

Imagine quantizing a Gaussian with variance σ^2 using only one bit. The optimal threshold (so our quantization is $1\{|X-\mu| \leq t\}$) is $t = \sqrt{2/\pi}\sigma$ and the mean squared error is $E(X - \hat{X}(X))^2 = \frac{\pi-2}{\pi}\sigma^2 \approx 0.36\sigma^2$.

Let's take $X \sim N(0, \sigma^2)$ and compute $R(D)$.

$$R(D) = \min_{p_{\hat{X}|X}, E(X-\hat{X})^2 \leq D} I(X;\hat{X})$$

We have

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X \mid \hat{X}) \\ &= h(X) - h(X - \hat{X} \mid \hat{X}) \\ &\geq h(X) - h(X - \hat{X}) \\ &\geq \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e D \\ &= \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

since we have a bound on the second moment of $X - \hat{X}$.

Remember the Gaussian channel coding theorem: if $X \sim N(0,p)$ and $Z \sim N(0,N)$, then $Y = X+Z$ has $I(X;Y) = \frac{1}{2}\log(1+\frac{P}{N})$.

Thus consider a channel with $\hat{X} = X \sim N(0,\sigma^2 - D), Z \sim N(0,D)$. Then $X - \hat{X} = Z$ has the required moment bound and the result says that our lower bound is achievable, i.e. $I(X;\hat{X}) = \frac{1}{2}\log(\frac{\sigma^2}{D})$.

If we want to examine the distortion achieved by the one bit quantizer, look at

$$R(0.36\sigma^2) = \frac{1}{2}\log\frac{1}{0}.36 = 0.737 \text{ bits}$$

This says we could compress 100 symbols into about 74 bits and achieve the same error as the 1 bit quantizer would have with 100 bits.

"What the point is, is that you guys are supposed to be wowed."

"If X is a label, cats, trees, dogs, students, whatever... I didn't mean to rank you guys in that order."

Proof: achievability

A note: we have good channel codes, but we don't have constructions for good lossy compression schemes in general. The proof is, again, a random coding argument.

We want to proof that there exist codes operating at rate close to $R(D)$ that allow me to achieve distortion D.

Fix $P_{\hat{X}|X}$ such that $Ed(X,\hat{X}) \leq D$. We want to show that there exists a sequences of $(2^{nR}, n)$ codes that have rate $R \approx I(X;\hat{X})$ and achieve $E[d(X^n, g_n \circ f_n(X^n))] \leq D + \epsilon$ as $n \to \infty$.

Define a distortion typical set:

$$\begin{aligned} A_{d,\epsilon}^{(n)} = \{(x^n, \hat{x}^n) : \quad & \left|-\frac{1}{n}\log p_{X^n\hat{X}^n}(x^n,\hat{x}^n) - H(X,\hat{X})\right| < \epsilon \\ & \left|-\frac{1}{n}\log p_{X^n}(x^n) - H(X)\right| < \epsilon \\ & \left|-\frac{1}{n}\log p_{\hat{X}^n}(\hat{x}^n) - H(\hat{X})\right| < \epsilon \\ & |d(x^n,\hat{x}^n) - Ed(X,\hat{X})| < \epsilon\} \end{aligned}$$

This is just the typical set for $(X, \hat{X})$ intersected with the set

$$\{(x^n, \hat{x}^n) : |d(x^n, \hat{x}^n) - Ed(X, \hat{X})| < \epsilon\}$$

and by the weak law of large numbers,

$$d(X^n, \hat{X}^n) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i) \to Ed(X, \hat{X}) \quad \text{i.p.}$$

Claim 1: $P(A_{d,\epsilon}^{(n)}) \to 1$ as $n \to \infty$, shown above.

Claim 2:

$$p_{\hat{X}^n}(\hat{x}^n) \geq p_{\hat{X}^n|X^n}(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X})+3\epsilon)}$$

$\forall \hat{x}^n, x^n \in A_{d,\epsilon}^{(n)}$. Why?

$$p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = p(\hat{x}^n) \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \geq p(\hat{x}^n) \frac{2^{-n(H(X,\hat{X})-\epsilon)}}{2^{-n(H(X)-\epsilon)} 2^{-n(H(\hat{X})-\epsilon)}}$$

Rearrange to get the desired result, since the mutual information appears in the exponent.

Claim 3:

$$0 \leq x, y \leq 1 \quad \Rightarrow \quad (1 - xy)^n \leq 1 - x + e^{-yn}$$

Note that $(1 - xy)^n$ is convex and non-increasing in x for fixed y. $1 - x + e^{-yn}$ is linear in x and non-increasing for fixed y. Thus we only need to check if the inequality holds at the endpoints. This is trivial at $x = 0$ and follows from $1 - y \leq e^{-y}$ for $x = 1$.

"We're done with the setup, so now we get to the juicy bits. Where all the magic happens."

Random coding

Remember, we already fixed $P_{\hat{X}|X}$ such that $Ed(X, \hat{X}) \leq D$.

1) Generate 2^{nR} sequences $\hat{X}^n(i), i = 1, \ldots, 2^{nR}$ i.i.d. $\sim P_{\hat{X}}$.

2) Typical set encoding: for each X^n, select i such that $(X^n, \hat{X}^n(i)) \in A_{\epsilon,d}^{(n)}$ if possible.
   - If multiple i's, then break ties arbitrarily.
   - If no i's are found, then just send $i = 1$. This happens with probability p_e.

What is the distortion for this scheme? Remember that d is bounded by some d_{max}.

$$\begin{aligned} Ed(X^n, \hat{X}^n(i)) &\leq (1 - p_e)(D + \epsilon) + p_e d_{max} \\ &\leq D + \epsilon + p_e d_{max} \end{aligned}$$

Want to show: $p_e \to 0$ provided $R > I(X; \hat{X})$.

$$P((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = 1 - \sum_{\hat{x}^n} p_{\hat{X}^n}(\hat{x}^n) 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\}$$

$$P(!\exists i : (x^n, \hat{X}^n(i)) \in A_{d,\epsilon}^{(n)}) = \left(1 - \sum_{\hat{x}^n} p_{\hat{X}^n}(\hat{x}^n) 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\}\right)^{2^{nR}}$$

$$\begin{aligned}
p_e &= \sum_{x^n} p_{X^n}(x^n) \left(1 - \sum_{\hat{x}^n} p_{\hat{X}^n}(\hat{x}^n) 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\}\right)^{2^{nR}} \\
&\leq \sum_{x^n} p_{X^n}(x^n) \left(1 - \sum_{\hat{x}^n} p_{\hat{X}^n|X^n}(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\}\right)^{2^{nR}} \\
&= \sum_{x^n} p_{X^n}(x^n) \left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p_{\hat{X}^n|X^n}(\hat{x}^n|x^n) 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\}\right)^{2^{nR}} \\
&\leq \sum_{x^n} p_{X^n}(x^n) \left(1 - \sum_{\hat{x}^n} p_{\hat{X}^n|X^n}(\hat{x}^n|x^n) 1\{(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}\} + e^{-2^{-n(I(X;\hat{X})+3\epsilon)} 2^{nR}}\right) \quad \text{applying Claim 3} \\
&= P((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) + e^{-2^{n(R-(I(X;\hat{X})+3\epsilon))}}
\end{aligned}$$

The first term goes to zero and the second term goes to zero if $R > I(X;\hat{X}) + 3\epsilon$.

October 25

Review of rate-distortion theory

X^n ∼i.i.d. p_X.

$$X^n \longrightarrow \text{encoder} \overset{nR \text{ bits}}{\longrightarrow} \text{decoder} \longrightarrow \hat{X}^n$$

The expected distortion is $Ed(X^n, \hat{X}^n) \leq D$. We care about the function $R(D)$: the minimum rate needed to send with expected distortion at most D. This is an “operational” definition, a statement about schemes and their possible performance (rather than some formula).

Theorem:

$$R(D) = \min_{\substack{p_{\hat{X}|X}: \\ Ed(X,\hat{X}) \leq D}} I(X;\hat{X})$$

Last time, we proved “achievability”: if $R > R(D)$, there exists a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} Ed(X^n, \hat{X}^n) \leq D$. Proof idea: random coding argument.

Informal review of big ideas in achievability proof

First, we picked a channel that satisfied the constraint: $P_{\hat{X}|X}$ such that $Ed(X,\hat{X}) \leq D$ (remember, P_X, the distribution of the source, is fixed).

Then we pick 2^{nR} sequences out of $\hat{\mathcal{X}}^n$: $\hat{X}^N(i)$ at random for $i = 1, \ldots, 2^{nR}$.

We look at the “distortion balls” in $\mathcal{X}^n$: sets of all sequences X^n with $d(X^n, \hat{X}^n(i)) \leq D$.

If you fall within the ball associated with index i, the bit string you send is just i.

Our long string of inequalities was just trying to show that if you picked enough $\hat{X}^N(i)$ sequences (i.e., if R was big enough), then the union of all of the distortion balls would be so big that the probability of observing an X^n sequence that *wasn't* in one of the balls would be tiny.

Plunk down enough balls so that the probability the encoder screws up is negligible.

Converse

Key ingredients: data processing inequality and Jensen's inequality.

Lemma

$R(D)$ is a convex function of D, as in the function

$$R(D) = \min_{\substack{p_{\hat{X}|X}: \\ Ed(X,\hat{X}) \leq D}} I(X;\hat{X}),$$

not the operational definition! Proof: let $P^{(0)}_{\hat{X}|X}$ achieve $R(D_0)$ and $P^{(1)}_{\hat{X}|X}$ achieve $R(D_1)$, then define

$$P^{(\lambda)}_{\hat{X}|X} = \lambda P^{(0)}_{\hat{X}|X} + \bar{\lambda} P^{(1)}_{\hat{X}|X}$$

where $\lambda + \bar{\lambda} = 1$.

Since mutual information is convex in $P_{\hat{X}|X}$ for fixed P_X, we have

$$I_{P^{(\lambda)}_{\hat{X}|X}}(X;\hat{X}) \leq \lambda I_{P^{(0)}_{\hat{X}|X}}(X;\hat{X}) + \bar{\lambda} I_{P^{(1)}_{\hat{X}|X}}(X;\hat{X}) = \lambda R(D_0) + \bar{\lambda} R(D_1).$$

Now,

$$E_{P^{(\lambda)}_{\hat{X}|X}} d(X,\hat{X}) = \lambda E_{P^{(0)}_{\hat{X}|X}} d(X,\hat{X}) + \bar{\lambda} E_{P^{(1)}_{\hat{X}|X}} d(X,\hat{X}) \leq \lambda D_0 + \bar{\lambda} D_1.$$

Together, these imply that

$$\lambda R(D_0) + \bar{\lambda} R(D_1) \geq I_{P^{(\lambda)}_{\hat{X}|X}}(X;\hat{X}) \geq \min_{\substack{p_{\hat{X}|X}: \\ Ed(X,\hat{X}) \leq \lambda D_0 + \bar{\lambda} D_1}} I(X;\hat{X}) = R(\lambda D_0 + \bar{\lambda} D_1)$$

as desired.

"Second board, I will leave unerased for now. Better write fast!"

Proof of converse

Consider a $(2^{nR}, n)$ code with encoder f_n and decoder g_n that achieves $Ed(X^n, \hat{X}^n) = D$, with $\hat{X}^n = g_n(f_n(X^n))$.

We want to show that $R \geq R(D)$.

$$
\begin{aligned}
nR &\geq H(\hat{X}^n) \\
&\geq H(\hat{X}^n) - H(\hat{X}^n \mid X^n) \\
&= I(\hat{X}^n; X) \\
&= H(X^n) - H(X^n \mid \hat{X}^n) \\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \hat{X}^n, X_1, \ldots, X_{i-1}) \\
&\geq \sum_{i=1}^{n} (H(X_i) - H(X_i \mid \hat{X}_i)) \\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\
&\geq \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) \quad \text{by the defn. of } R(\cdot) \\
&\geq nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i)\right) \quad \text{by our convexity lemma} \\
&= nR(D) \quad \text{by the defn. of how we apply } d \text{ to a vector}
\end{aligned}
$$

“Gosh, I love converse arguments.”

Comments on tightness of inequalities

$nR \geq H(\hat{X}^n)$:

this is tight when all nR reproductions are equally likely.

$H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n \mid X^n)$:

this is tight when $\hat{X}^n$ is a deterministic function of X^n.

$\sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \hat{X}^n, X_1, \ldots, X_{i-1}) \geq \sum_{i=1}^{n}(H(X_i) - H(X_i \mid \hat{X}_i))$:

this is tight when only $\hat{X}_i$ is useful for reproducing X_i.

$\sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i))$:

this is tight when $P_{\hat{X}_i|X_i}$ is the minimizing conditional distribution from the definition of R.

$\sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) \geq nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i)\right)$:

this is tight if either R is linear or all the $Ed(X_i, \hat{X}_i) = D$ (so maximal distortion is used for reconstructing each symbol).

Why this is class is different: “rarely do you have derivations that are as... long”.

Joint source channel coding

"I want to post this picture on Instagram, so the whole *world* can enjoy that I was at the beach."

We have $V^m \sim$ i.i.d. p_V.

$$V^m \longrightarrow \text{encoder} \xrightarrow{X^n} P_{Y|X} \xrightarrow{Y^n} \text{decoder} \longrightarrow \hat{V}^m$$

I observed something, I want to communicate it, nature is corrupting the communication, and at the other end, someone recovers what I see ($Ed(V^m, \hat{V}^m) \leq D$).

Theorem: Distortion D is achievable if and only if $R(D) < BC$ (for $B = \frac{n}{m}$, the bandwidth mismatch).

"What this result says, is that there is no magic to be done."

This result says that separation is optimal.

We should use a rate distortion source code at rate $R(D)$ to compress into an index i using $mR(D)$ bits. Now we use a channel code at rate C, send the result through the channel, and decode an estimate $\hat{i}$, which is given to the R-D decoder.

Thus optimal schemes for compression and communication through a channel *do not need to depend on the content of the message.* The two tasks can be separated without any loss of optimality.

$$nC \geq I(X^n; Y^n) \geq I(V^m; \hat{V}^m) \geq mR(D)$$

can be demonstrated by repeating our arguments for source coding and channel coding proofs. The middle step comes from the data processing inequality.

October 27

Today: approximating probability distributions with dependence trees (Chow-Liu, 1968).

Dependence trees

$P(x)$ is a joint distribution on n variables, so $x = (x_1, \ldots, x_n)$.

Our goal is to approximate $P(x)$ by a "second-order" distribution.

$m_1, \ldots,_n$ is a permutation of $1, 2, \ldots, n$. We can factor

$$P(x) = \prod_{i=1}^{n} P(x_{m_i} \mid x_{m_1}, \ldots, x_{m_{i-1}})$$

But estimating these conditional distributions can be quite difficult because of the curse of dimensionality.

A second order or "tree-dependence" distribution has this form:

$$P_t(x) = \prod_{i=1}^{n} P(x_{m_i} \mid x_{m_{j(i)}})$$

It is second order because each factor has two variables in it. We can draw this as a tree where $m_{j(i)}$ is the parent of m_i in the tree.

Finding the best second order approximation

Optimization problem:

$$\min_{t \in T_n} D(P||P_t)$$

where T_n = set of trees on n vertices. Note that the set of trees with n vertices has size n^{n-2}! However, using KL divergence as a distance leads to a clean solution (unlike something such as total variation). Here, we are assuming we are given P and are only optimizing over the set of trees.

For a tree $t \in T_n$, let $j(i)$ be the parent of i.

Definition: A maximum weight dependence tree is a tree t such that

$$\sum_{i=1}^{n} I(X_i; X_{j(i)}) \geq \sum_{i=1}^{n} I(X_i; X_{j'(i)}), \quad \forall t' \in T_n.$$

This is the minimum spanning tree over all pairwise mutual informations, so it can be computed.

Theorem: $t^* \in \operatorname{argmin}_t D(P||P_t)$ if and only if it is a maximum weight dependence tree.

"The proof is remarkably *cute*."

Proof:

$$\begin{aligned}
D(P||P_t) &= \sum_x p(x) \log \frac{p(x)}{p_t(x)} \\
&= \sum_x p(x) \log p(x) - \sum p(x) \sum_{i=1}^{n} \log p(x_i \mid x_{j(i)}) \\
&= -H(X) - \sum_x p(x) \sum_{i=1}^{n} \log \frac{p(x_i, x_{j(i)})}{p(x_{j(i)}) p(x_i)} - \sum_x p(x) \sum_{i=1}^{n} \log p(x_i) \\
&= -H(X) + \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} I(X_i; X_{j(i)})
\end{aligned}$$

At this point, we've already seen the magic, since only the third term depends on the tree. Thus

$$\min_t D(P||P_t) \iff \max_t \sum_{i=1}^{n} I(X_i; X_{j(i)})$$

as desired.

Now we've reduced to a much simpler estimation problem than estimating the whole distribution: estimating the pairwise mutual information. In this case, error decays with the number of samples collected and does not depend on the dimension at all.

Suppose we have independent n dimensional samples. If we just use the plug-in empirical distribution estimator for mutual information, the best tree we get will be the maximum likelihood estimator of the real best tree. This is shown in the appendix of the paper.

Say I have data $X^{(1)}, \dots, X^{(n)}$ (where you have n samples, n is no longer the dimension of the data). I want to put it through an estimator and get an estimate of the dependence tree.

Inside our estimator box, we could estimate all the mutual informations, compute the max weight tree, and return the estimate of the dependence tree.

How do we estimate mutual information? Since

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

what we're more concerned with is estimating entropies.

Estimating entropy

How should I estimate $H(P)$ given n i.i.d. samples from P?

These results are from the last couple years, so we'll omit the proofs.

Classical statistics

Consider n samples drawn from an alphabet with $|\mathcal{X}| = S$, S fixed. We want to find an optimal estimator of $H(P)$ for $n \to \infty$.

In this case, the empirical entropy $H(P_n)$ (with P_n = empirical distribution), being

$$H(P_n) = \sum \hat{p}_i \log \frac{1}{\hat{p}_i}$$

is the plug-in estimator.

Claim: $H(P_n)$ is MLE. Also, $H(P_n)$ is asymptotically efficient (saturates the Cramer-Rao bound).

The problem with this is that... n is going to infinity.

Question: what about non-asymptotics?

"Is this purely an academic question? Do we care about *finite* n? Surely not, it seems in this class. But data costs money."

What if n is not huge compared to S? This is the usual setting nowadays.

Decision theoretic framework

For an estimate $\hat{H}_n$, the risk is

$$R_n^{max}(\hat{H}_n) = \sup_{P \in \mathcal{M}_s} E_P(H(P) - \hat{H}_n)^2$$

The minimax risk is

$$\inf_{\hat{H}_n} \sup_{P \in \mathcal{M}_s} E_P(H(P) - \hat{H}_n)^2$$

Classical asymptotic says

$$E_P(H(P) - H(P_n))^2 \sim \frac{\mathrm{Var}(-\log P(X))}{n}$$

as n grows large.

Now,

$$\sup_{P \in \mathcal{M}_S} \mathrm{Var}(-\log P(X)) \leq \frac{3}{4}(\log S)^2$$

so does $n >> (\log S)^2$ imply consistency? Turns out it doesn't.

$$\text{Risk} = E_P(H(P) - H(P_n))^2 = (E_P(\hat{H}) - H(P))^2 + \mathrm{Var}_P \hat{H} = bias^2 + variance.$$

From a 2014 paper,

$$R_n^{max}(H(P_n)) \approx \frac{S^2}{n^2} + \frac{(\log S)^2}{n}$$

up to constants. The first term comes from bias, the second from variance. So when n is comparable to S, the bias term becomes important.

We can do significantly better... Next time.

November 1

Computation of $R(D)$ and C

Alternating minimization algorithm: the basic idea. To minimize $\|a - b\|^2$ over two sets A, B, alternately minimize in each coordinate.

Generally, we can minimize distance functions on two convex sets, provided the distance function is well behaved.

Relative entropy is a good distance function.

Let $A = \{Q_{X\hat{X}} : E_{Q_{X\hat{X}}} d(x, \hat{x}) \leq D, \sum_{\hat{x}} Q_{X\hat{X}}(x, \hat{x}) = P_X(x)\}$. (In rate distortion, the source distribution is fixed and what can change is the channel.) This is a convex set of distributions since the constraints are linear in Q.

We want to compute

$$R(D) = \min_{Q_{X\hat{X}} \in A} I(X; \hat{X}) = \min_{Q \in A} D(Q_{X\hat{X}} || Q_X Q_{\hat{X}}) = \min_{Q \in A} D(Q_{X\hat{X}} || P_X Q_{\hat{X}})$$

Lemma: $R(D) = \min_{Q_{X\hat{X}} \in A} \min_{R_{\hat{X}}} D(Q_{X\hat{X}} || P_X R_{\hat{X}})$.

$$D(Q_{X\hat{X}} || P_X R_{\hat{X}}) - D(Q_{X\hat{X}} || P_X Q_{\hat{X}}) = D(Q_{\hat{X}} || R_{\hat{X}}) \geq 0$$

So to implement alternating minimization algorithm, we need to solve two problems:

1. Given $Q_{X\hat{X}}$, find $R_{\hat{X}}$ that minimizes $D(Q_{X\hat{X}} || P_X R_{\hat{X}})$. Our lemma implies that $R^*_{\hat{X}} = Q_{\hat{X}}$.
2. Given $R_{\hat{X}}$, find $Q_{X\hat{X}} \in A$ that minimizes $D(Q_{X\hat{X}} || P_X R_{\hat{X}})$. Lemma 2:

$$Q_{\hat{X}|X}(\hat{x}|x) = \frac{R_{\hat{X}}(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} R_{\hat{X}}(\hat{x}') e^{-\lambda d(x,\hat{x}')}}$$

   λ is chosen so that $E_{Q_{X\hat{X}}} d(X, \hat{X}) = D$.

   Proof: Lagrange multipliers. We have

$$J(Q_{\hat{X}|X}) = D(Q_{\hat{X}|X} P_X || P_X R_{\hat{X}}) + \lambda_1 E_{Q_{X\hat{X}}} d(X, \hat{X}) + \lambda_2 \sum_{x,\hat{x}} P_X(x) Q_{\hat{X}|X}(\hat{x}|x)$$

   Differentiate w.r.t. the conditional distribution to get the multipliers.

November 3

Information Theory and Statistics

Let $X_1, \ldots, X_n$ be a sequence of symbols from $\mathcal{X} = \{a_1, \ldots, a_n\}$.

Definition: the *type* of X^n, $\mathbb{P}_{X^n}$ is the empirical distribution associated to $X^n = (X_1, \ldots, X_n)$,

$$\mathbb{P}_X(x) = \frac{1}{n} \#\{i : x_i = x\} = \text{ empirical frequency of } x \text{ in } x^n$$

Definition: $\mathcal{P}_n$ is the set of types with denominator n.

Example: If $\mathcal{X} = \{0, 1\}$,

$$\mathcal{P}_n = \{(0/n, n/n), (1/n, (n-1)/n), \ldots, (n/n, 0/n)\}$$

The type class of $P \in \mathcal{P}_n$, denoted by $T(P)$, is

$$T(P) = \{x^n \in \mathcal{X}^n : \mathbb{P}_{X^n} = P\}$$

Example: If $P = (3/8, 5/8)$, $\mathcal{X} = \{0, 1\}$, then $T(P) =$ all $\binom{8}{3}$ binary vectors with exactly 3 zeros.

Theorem: $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$.

If $P \in \mathcal{P}_n$, then

$$P = \left(\frac{n_1}{n}, \ldots, \frac{n_{|\mathcal{X}|}}{n}\right), \quad \sum_{i=1}^{n} n_i = n$$

This is a very crude upper bound, but the point is that the number of type classes (or number of types) is polynomial in n.

Theorem: Let $X_1, \ldots, X_n$ be $\sim$ i.i.d. Q. Then the probability of X^n is

$$Q^n(X^n) = 2^{-n(H(\mathbb{P}_{X^n})+D(\mathbb{P}_{X^n}||Q))}$$

Note that the right hand side only depends on the type and Q.

Also, recall that $Q_{X^n}(X^n) \approx 2^{-nH(Q)}$ for typical X^n.

Proof:

$$\begin{aligned} 2^{-n(H(\mathbb{P}_{X^n})+D(\mathbb{P}_{X^n}||Q))} &= 2^{n\left(\sum_{a\in X} \mathbb{P}_{X^n}(a)\log Q(a)\right)} \\ &= \prod_{a\in\mathcal{X}} Q(a)^{n\mathbb{P}_{X^n}(a)} \\ &= Q^n(X^n) \end{aligned}$$

since $n\mathbb{P}_{X^n}(a)$ is the number of times that a appears in X^n.

"It's like a great symphony, in four acts. Oh by the way, did anyone ever figure out what it is, about this class..."

Theorem:

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

i.e., type class for P has roughly $2^{nH(P)}$ sequences.

Proof:

$$\begin{aligned} 1 &\geq P^n(T(P)) \\ &= \sum_{x^n\in T(P)} P^n(x^n) \\ &= |T(P)|2^{-nH(P)} \end{aligned}$$

Just to verify the last step,

$$2^{-nH(P)} = 2^{n\sum p(a)\log p(a)} = \prod p(a)^{np(a)}$$

For the lower bound, we will use the fact that

$$P^n(T(P)) \geq P^n(T(\hat{P})) \quad \forall \hat{P} \in P_n$$

Then

$$\begin{aligned} 1 = \sum_{Q\in P_n} P^n(T(Q)) \leq \sum_{Q\in P_n} \max_Q P^n(T(Q)) \leq |P_n|P^n(T(P)) \\ \leq (n+1)^{|\mathcal{X}|}P^n(T(P)) = (n+1)^{|\mathcal{X}|}|T(P)|2^{-nH(P)} \end{aligned}$$

To prove the fact we assumed:

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(\hat{P}))} &= \frac{T(P)\prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{T(\hat{P})\prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \\
&= \frac{\binom{n}{nP(a_1),nP(a_2),\ldots,nP(a_n)}}{\binom{n}{n\hat{P}(a_1),n\hat{P}(a_2),\ldots,n\hat{P}(a_n)}} \cdot \prod_{a\in\mathcal{X}} P(a)^{nP(a)-\hat{P}(a)} \\
&= \frac{(nP(a_1))!\cdots(nP(a_n))!}{(n\hat{P}(a_1))!\cdots(n\hat{P}(a_n))!} \cdot \prod_{a\in\mathcal{X}} P(a)^{nP(a)-\hat{P}(a)} \\
&= \prod_{a\in\mathcal{X}} \frac{(nP(a))!}{(n\hat{P}(a))!} P(a)^{nP(a)-\hat{P}(a)} \\
&\geq \prod_{a\in\mathcal{X}} (nP(a))^{n\hat{P}(a)-nP(a)} P(a)^{nP(a)-n\hat{P}(a)}, \quad \text{using ineq. } \frac{m!}{n!} \geq n^{m-n} \\
&= \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} \\
&= n^{n\sum_a(\hat{P}(a)-P(a))} \\
&= 1 \quad \text{“and holy crap, we get 1”}
\end{aligned}$$

“You begin to question whether you’re going in the right direction, or whether you should really be doing this at all.”

“Second year is the worst, in my opinion, because you start working on problems, and everything you find out has already been done before.”

Theorem: For any $P \in P_n$ and any distribution Q, the probability of type class $T(P)$ under Q is

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

Proof:

$$\begin{aligned}
Q^n(T(P)) &= \sum_{x^n\in T(P)} Q^n(x^n) \\
&= \sum_{x^n\in T(P)} 2^{-n(D(P||Q)+H(P))} \\
&= |T(P)| 2^{-n(D(P||Q)+H(P))}
\end{aligned}$$

then use the previous bound.

The four main things: summary

1. $|P_n| \leq (n+1)^{|\mathcal{X}|}$: the number of types is polynomial, so it’s much much smaller than the total number of sequences.
2. $Q^n(x^n) = 2^{-n(H(P)+D(P||Q))}$ when $x^n \in T(P)$. Probability is approximately entropy but also plus a correction factor.

3. $|T(P)| \doteq 2^{nH(P)}$ for $P \in P_n$.
4. $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$ for $P \in P_n$, basically a combination of 3 and 4. More exactly, this says

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

If you connect back to typical sets, they make sense (except the first one).

A first application

"Now that you were so patient and so nice, and laughed at all my jokes, we finally get to do something interesting."

Theorem: Let $X_1, \dots, X_n$ be i.i.d. $\sim P$. Then

$$P(D(\mathbb{P}_{X^n}||P) > \epsilon) \leq 2^{-n(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n})}$$

This is cool because then Borel-Cantelli implies that

$$D(\mathbb{P}_{X^n}||P) \to 0$$

almost surely.

November 8

Reference 25 in the book for hw 6 prob 2.

Midterm 2 focused on stuff after midterm 1: channel capacity, rate distortion, connections to statistics (type), estimation of information measures (chow liu type question showing opt problem has a nice solution).

"The material is really beautiful."

"This just got way too deep for me. Boxes and cats, I don't know."

Laws of large numbers

Given $\epsilon > 0$ and a distribution Q, define

$$T_Q^\epsilon = \{x^n : D(\mathbb{P}_{X^n}||Q) \leq \epsilon\}$$

Then

$$\begin{aligned} 1 - Q^n(T_Q^\epsilon) &= \sum_{P: D(P||Q) > \epsilon, P \in \mathcal{P}_n} Q^n(T(P)) \\ &\leq \sum_{P: D(P||Q) > \epsilon, P \in \mathcal{P}_n} 2^{-nD(P||Q)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \end{aligned}$$

This is the probability of observing an X^n sampled from Q such that $D(\mathbb{P}_{X^n}||Q) > \epsilon$. This is the law of large numbers we wanted to prove at the end of last time. We applied points 1 and 4 from last time's summary.

As mentioned last time, this means that $D(\mathbb{P}_{X^n}||Q) \to 0$ with probability 1, by Borel-Cantelli.

Large deviations

Consider $X_1, X_2, \ldots$ i.i.d. $\sim Q$ (finite alphabet $\mathcal{X}$).

We know that the WLLN tells us

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > EX_1 + \epsilon\right) \to 0 \quad \text{as } n \to \infty$$

However, we don't know *how* this probability is decaying. Chebyshev says it's decaying as $\frac{\text{Var}(X)}{\epsilon^2 n^2}$ (maybe?). Basically, polynomially fast in n. Can we say anything stronger?

We can also view this as the probability that $\sum_{i=1}^{n} X_i$ deviates from its expectation by $n\epsilon$, a large deviation (a constant fraction of its mean). It turns out that

$$P\left(\sum_{i=1}^{n} X_i \geq nEX_1 + n\epsilon\right) \doteq 2^{-nE}$$

where E is an exponent that we can compute explicitly (the probabilities agree to first order in the exponent).

Example. Let $X_i = 1$ with probability p, so $Q(1) = p$ and $Q(0) = 1 - p$.

$$P\left(\sum X_i \geq n(p+\epsilon)\right) = \sum_{P \in \mathcal{P}_n : P(1) \geq p+\epsilon} Q^n(T(P))$$

We can bound this some above by

$$|\mathcal{P}_n| 2^{-n \min_{P \in \mathcal{P}_n : P(1) \geq p+\epsilon} D(P||Q)}$$

and below by

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n \min_{P \in \mathcal{P}_n : P(1) \geq p+\epsilon} D(P||Q)}$$

which gives that exponent we wanted.

Sanov's theorem

Let $X_1, X_2, \ldots$ be i.i.d. $\sim Q$. Let $E \subseteq P(\mathcal{X})$ (probability distributions on $\mathcal{X}$) be a collection of distributions. Then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}$$

where $P^* = \text{argmin}_{P \in E} D(P||Q)$. Moreover, if E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q)$$

(i.e., lower bound matches upper bound).

Proof:

$$Q^n(E) = \sum_{P\in E\cap\mathcal{P}_n} Q^n(T(P)) \le \sum_{P\in E\cap\mathcal{P}_n} 2^{-nD(P||Q)} \le (n{+}1)^{|\mathcal{X}|} 2^{-n\min_{P\in E\cap\mathcal{P}_n} D(P||Q)} \le (n{+}1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}$$

That's the first part proved. For the lower bound

$$\begin{aligned} Q^n(E) &= \sum_{P\in E\cap\mathcal{P}_n} Q^n(T(P)) \\ &\ge Q^n(T(P_n)) \quad \text{for any } P_n \in \mathcal{P}_n \cap E \\ &\ge \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)} \end{aligned}$$

We need to find the sequence $\{P_n : P_n \in E \cap \mathcal{P}_n\}_{n\ge1}$ such that $D(P_n||Q) \to D(P^*||Q)$.

If E has a nonempty interior and E is the closure of that interior, then this is possible.

November 10

Review of Sanov's theorem

Statement: Let $X_1, X_2, \dots$ be i.i.d. $\sim Q$. Let $E \subseteq P(\mathcal{X})$ (probability distributions on $\mathcal{X}$) be a collection of distributions. Then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \le (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}$$

where $P^* = \operatorname{argmin}_{P\in E} D(P||Q)$. Moreover, if E is the closure of its interior, then

$$\frac{1}{n}\log Q^n(E) \to -D(P^*||Q)$$

(i.e., lower bound matches upper bound).

Notes: The overall space is probability distributions on $\mathcal{X}$. Q is some point in this space, and E is some subset of this space. E is basically a set of properties of sequences that we may observe. So we are asking, what is the probability that I observe a sequence with the property E? Sanov's tells us that this probability is exponentially small in the distance from Q to the set E, where distance is measure in KL divergence.

Example: Suppose we have a fair coin. What is the probability of at least 700 heads in 1000 tosses? We have

$$E = \{P : \sum_{x\in 0,1} P(x)\cdot x \ge .7\}$$

Sanov's theorem tells us that

$$\frac{1}{n}\log P\{\ge 700 \text{ heads}\} \approx -D((0.7, 0.3)||(0.5, 0.5)) = 0.119$$

The bound sandwich is large for 1000, but quite tight for 10000.

“What’s the best application of this theorem?” “Oh, you’re gonna see an *awesome* application of this theorem.”

All of this has been for discrete alphabets. There is a version for continuous spaces.

$$-\inf_{P\in int(E)} D(P||Q) \leq \liminf \frac{1}{n}\log Q^n(E) \leq \limsup \frac{1}{n}\log Q^n(E) \leq -\inf_{P\in cl(E)} D(P||Q)$$

So if the left and right optimization problems have answers that coincide, the limit exists and it nails the exponential decay of Q^n.

“This is like, Adult Sanov. Wait, no. Call it, Sanov for Pros.”

It’s hard to distill the non-asymptotic bounds since they depend on the cardinality of the alphabet.

Conditional limit theorem

Suppose I am manufacturing bolts. Each is supposed to nominally weigh 10 grams. Now, I find a batch of 1000 bolts weighs more than 10.5 kilograms. What is the probability any given bolt weighs, say, 11 grams?

Suppose $X_1, X_2, \ldots$ are i.i.d. $\sim Q$, and we observe $\mathbb{P}_{X^n} \in E$, $E \not\ni Q$, and is closed, convex set.

$$P(X_1 = a \mid \mathbb{P}_{X^n} \in E) \to P^*(a)$$

with high probability, where

$$P^* = \operatorname*{argmin}_{P\in E} D(P||Q)$$

Pythagorean theorem for relative entropy

For $E \subset P(\mathcal{X})$ closed and convex and $Q \notin E$, define $P^* = \operatorname{argmin}_{P\in E} D(P||Q)$.

$$D(P||Q) \geq D(P||P^*) + D(P^*||Q)$$

for all $P \in E$.

Proof: Let $P \in E$ and define $P_\lambda = \lambda P + \bar{\lambda}P^*$ for $\lambda \in [0,1]$. The clever part: by defintion of P^*,

$$\frac{d}{d\lambda} D(P_\lambda||Q) \geq 0$$

at $\lambda = 0$. This is where all the magic happens, now it’s just a matter of calculation

$$\begin{aligned}\frac{d}{d\lambda} D(P_\lambda||Q) &= \frac{d}{d\lambda}\sum_x P_\lambda(x)\log\frac{P_\lambda(x)}{Q(x)}\\ &= \sum_x \left[(P(x) - P^*(x))\log\frac{P_\lambda(x)}{Q(x)} + (P(x) - P^*(x))\right]\end{aligned}$$

The second term in the sum cancels since both distributions sum to 1.

At $\lambda = 0$, since $P_0 = P^*$, we have

$$\begin{aligned}
0 &\leq \frac{d}{d\lambda}|_{\lambda=0} D(P_\lambda||Q) \\
&= \sum_x (P(x) - P^*(x)) \log \frac{P_0(x)}{Q(x)} \\
&= \sum_x P(x) \log \frac{P^*(x)P(x)}{Q(x)P(x)} - D(P^*||Q) \\
&= D(P||Q) - D(P||P^*) - D(P^*||Q)
\end{aligned}$$

Pinsker's inequality

$$D(P||Q) \geq \frac{\log e}{2} \|P - Q\|_1^2$$

For $A = \{x : P(x) \geq Q(x)\}$:

$$\begin{aligned}
\|P - Q\|_1 &= \sum_x |P(x) - Q(x)| \\
&= \sum_{x \in A} (P(x) - Q(x)) - \sum_{x \notin A} (P(x) - Q(x)) \\
&= P(A) - Q(A) - (1 - P(A) - (1 - Q(A)) \\
&= 2(P(A) - Q(A)) \\
&= \max_{B \in \mathcal{X}} 2(P(B) - Q(B))
\end{aligned}$$

For binary distributions, we get

$$p \log \frac{p}{q} + \bar{p} \log \frac{\bar{p}}{\bar{q}} \geq \frac{\log e}{2} (2(p - q))^2$$

(as an exercise).

Data processing for relative entropy: if we put a distribution P through a channel $P_{Y|X}$ and get P', and put Q through to get Q', then

$$D(P||Q) \geq D(P'||Q')$$

Define a channel such that $Y = 1\{X \in A\}$ where A is the same as before. Now by the DPI,

$$D(P||Q) \geq D((P(A), 1 - P(A))||(Q(A), 1 - Q(A))) \geq \frac{\log e}{2} (2(P(A) - Q(A)))^2 = \frac{\log e}{2} \|P - Q\|_1^2$$

So we turn the distributions into binary and then use our special case.

November 15

Midterm notes: probably 4 questions (differential entropy, rate distortion, connections to statistics (method of types), channel capacity, entropy power inequality (not proof though),. . .).

Review of last time

Conditional limit theorem: Let E be a closed and convex set of distributions on $\mathcal{X}$. $X_1, X_2, \ldots$ i.i.d. $\sim Q$. Then

$$P(X_1 = a \mid \mathbb{P}_{X^n} \in E) \to P^*(a)$$

in probability, where $P^* = \operatorname{argmin}_{P \in E} D(P||Q)$.

Some comments:

1. $P(X_1 = a \mid \mathbb{P}_{X^n} \in E)$ is a random variable, which is why we specify a type of convergence.
2. E is some property. For example: the sample mean exceeds some threshold.
3. Given that property, we see that the conditional probabilities behave according to P^*, i.e., they are very predictable and quantifiable.

Results we needed:

1. For $E \subset P(\mathcal{X})$ (distributions on $\mathcal{X}$) closed, convex, and $Q \notin E$,

   $$D(P||Q) \geq D(P||P^*) + D(P^*||Q)$$

   (triangle-like inequality). So KL divergence behaves like squared Euclidean distance.
2. Pinsker's inequality:

   $$D(P||Q) \geq \frac{\log e}{2} \|P - Q\|_1^2$$

   Another example of divergence behaving sort of like a squared norm.

Now we can actually prove the theorem.

Proof of conditional limit theorem

Define sets $S_t = \{P \in \mathcal{P}(\mathcal{X}) : D(P||Q) \leq t\}$. This set is convex.

Let $D^* = D(P^*||Q) = \min_{P \in E} D(P||Q)$.

We'll look at a neighborhood of the distribution P^*: $A = S_{D^*+2\delta} \cap E$ and $B = E \setminus S_{D^*+2\delta}$.

$$Q^n(B) = \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} Q^n(T(P)) \leq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} 2^{-nD(P||Q)} \leq (n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}$$

$$Q^n(A) \geq Q^n(S_{D^*+\delta} \cap E) = \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) \leq D^*+\delta} Q^n(T(P)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}$$

$$P(\mathbb{P}_{X^n} \in B \mid \mathbb{P}_{X^n} \in E) = \frac{Q^n(B \cap E)}{Q^n(E)} \leq \frac{Q^n(B)}{Q^n(A)} \leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}} = (n+1)^{2|\mathcal{X}|} 2^{-n\delta}$$

Thus

$$P(\mathbb{P}_{X^n} \in B \mid \mathbb{P}_{X^n} \in E) \to 0$$

This means that P^* overwhelmingly contains the probability in E, since everything else has exponentially smaller probability.

“The big picture is, oh wow that’s good. The big picture is literally this big picture here. I didn’t even plan that.”

If we flip things around, we see that

$$P(\mathbb{P}_{X^n} \in A \mid \mathbb{P}_{X^n} \in E) \to 1$$

as $n \to \infty$.

Now, for all $P \in A$, our fake triangle inequality says that

$$D(P||P^*) + D(P^*||Q) \leq D(P||Q) \leq D^* + 2\delta$$

Thus $D(P||P^*) \leq 2\delta$.

Therefore the event $\{\mathbb{P}_{X^n} \in A\}$ is contained in $\{D(\mathbb{P}_{X^n}||P^*) \leq 2\delta\}$. Thus

$$P(D(\mathbb{P}_{X^n}||P^*) \leq 2\delta \mid \mathbb{P}_{X^n} \in E) \to 1$$

as $n \to \infty$, since $P(\mathbb{P}_{X^n} \in A \mid \mathbb{P}_{X^n} \in E) \to 1$.

Thus

$$P(\|P - P^*\|_1 \leq \delta' \mid \mathbb{P}_{X^n} \in E) \to 1$$

So for any ϵ,

$$P(|P(a) - P^*(a)| \leq \epsilon \mid \mathbb{P}_{X^n} \in E) \to 1$$

which means $P(X_1 = a \mid \mathbb{P}_{X^n} \in E) \to P^*(a)$ in probability.

Fisher information and the Cramer-Rao bound

We have an indexed family of densities $f(x;\theta)$. For example,

$$f(x;\theta) = f(x - \theta)$$

where the mean is unknown.

An estimator for θ from sample size n is a function $T : \mathcal{X}^n \to \Theta$. The error of the estimator is $T(X^n) - \theta$, which is a random variable.

Example: $X_n \sim N(\theta, 1)$ i.i.d. An estimator for θ would be the sample mean.

An estimator is unbiased if

$$E_\theta T(X) = \int T(x) f(x;\theta) dx = \theta$$

We’d like to look at the error of unbiased estimators.

Preview: the Cramer-Rao bound tells us that the estimation error of an unbiased estimator is lower bounded by $1/J(\theta)$, where J is the Fisher information.

Example: if $f(x;\theta) = f_X(x - \theta)$, then

$$J(\theta) = \int \frac{f_X'(x)^2}{f_X(x)} dx$$

which looks like a measure of smoothness of the function f_X.