Information theory

Billy Fang
Instructor: Thomas Courtade

Fall 2016

These are my personal notes from an information theory course taught by Prof. Thomas Courtade. Most of the material is from [1]. Any errors are mine.

Contents

1 Entropy

[Absent from first lecture; refer to Chapter 2 of [1] for missing introductory material.]

$$H(X) = -\sum_x p(x) \log p(x) = -\mathbb{E} \log p(X).$$

$$H(Y \mid X) = \sum_x p(x) H(Y \mid X = x) = -\mathbb{E} \log p(Y \mid X).$$

Chain rule: $H(X, Y) = H(X) + H(Y \mid X)$.

$H(X \mid Y) \leq H(X)$ (conditioning reduces uncertainty), but it is not always true that $H(X \mid Y = y) \leq H(X)$ for each y (only true on average).

Pinsker's inequality: $D(p\|q) \geq \frac{1}{2 \ln 2} \|p - q\|_1^2$.

Entropy $H(p_1, \ldots, p_n)$ is concave in $(p_1, \ldots, p_n)$.

$H(X) \leq \log|\mathcal{X}|$. Proof: use Jensen's inequality. Let $p_X = (p_1, \ldots, p_n)$ and let $p_X^{(i)} := (p_i, \ldots, p_n, p_1, \ldots, p_{n-1})$. Then $H(p_X) = H(p_X^{(i)})$. So,

$$H(X) = \sum_{i=1}^n \frac{1}{n} H(p_X^{(i)}) \leq H\left(\frac{1}{n} \sum_{i=1}^n p_X^{(i)} \right) = H(1/n, \ldots, 1/n) = \log n.$$

$I(X; Y)$ is concave in p_X for fixed $p_{Y|X}$. It is convex in $p_{Y|X}$ for fixed p_X. To see the first one, note $I(X; Y) = H(Y) - H(Y \mid X)$. For fixed $p_{Y|X}$, we have $H(Y \mid X)$ is linear in p_X, and $H(Y)$ is concave function composed with a linear function $p_Y = \sum_x p_{Y|X}(y) p_X(x)$ of p_X.

Data processing inequality: if $X \to Y \to Z$ is a Markov chain (X and Z independent given Y), then $I(X; Z) \leq I(X; Y)$.

$$I(X; Z) \leq I(X; Z) + I(X; Y \mid Z) = I(X; Y, Z) = I(X; Y) + I(X; Z) = I(X; Y).$$

We discuss a connection between information and estimation. Consider the Markov chain $X \to Y \to \widehat{X}(Y)$ where Y is the noisy observation of X, and $\widehat{X}(Y)$ is an estimator of X based on Y. Let $P_e = \mathbb{P}(X \neq \widehat{X}(Y))$.

Theorem 1.1 (Fano's inequality)**.**

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

Proof. Let E be the indicator for $X \neq \widehat{X}$ so that $H(E) = H(P_e)$.

Using chain rule in two ways gives

$$\begin{aligned} H(E, X \mid \widehat{X}) &= H(X \mid \widehat{X}) + H(E \mid X, \widehat{X}) \\ &= H(X \mid \widehat{X}) \end{aligned}$$

and

$$\begin{aligned} H(E, X \mid \widehat{X}) &= H(E \mid \widehat{X}) + H(X \mid E, \widehat{X}) \\ &\leq H(E) + p_e H(X \mid \widehat{X}, E = 1) + (1 - p_e) H(X \mid \widehat{X}, E = 0) \\ &= H(P_e) + p_e \log|\mathcal{X}| \end{aligned}$$

□

The $\log|\mathcal{X}|$ can be replaced by $\log(|\mathcal{X}| - 1)$.

Corollary 1.2. With $P_e := \min_{\widehat{X}} \mathbb{P}(\widehat{X}(Y) \neq X)$, we have

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

Proof. By the data-processing inequality, $I(X;\widehat{X}) \le I(X;Y)$ which implies $H(X \mid \widehat{X}) \ge H(X \mid Y)$. □

We now clarify the interpretation of "entropy is the average number of bits needed to describe a random variable." Consider a function $f : \mathcal{X} \to \{0,1\}^*$ that maps the alphabet to arbitrary length bit strings. We want f to be injective (able to distinguish between letters in the alphabet). Let $\ell(f(x))$ be the length of this description of x. Then $\mathbb{E}[\ell(f(X))]$ is the average length of this description.

$$\begin{aligned}
\mathbb{E}[\ell(f(X))] &= \sum_x p(x)\ell(f(x)) \\
&= H(X) + \sum_x p(x) \log \frac{p(x)}{2^{-\ell(f(x))}} \\
&= H(X) - \sum_x p(x) \log \sum_{x'} 2^{-\ell(f(x'))} + \sum_x p(x) \log \frac{p(x)}{2^{-\ell(f(x))} / \sum_{x'} 2^{-\ell(f(x'))}} \\
&= H(X) - \log \sum_{x'} 2^{-\ell(f(x'))} + D(p_X \| Q) \\
&\ge H(X) - \log \sum_{x'} 2^{-\ell(f(x'))}.
\end{aligned}$$

Proposition 1.3.

$$\sum_{x'} 2^{-\ell(f(x'))} \le \log 2|\mathcal{X}|.$$

Proof. Consider maximizing the sum on the left-hand side. Respecting injectivity of f, this is the first $|\mathcal{X}|$ terms of the series

$$2^{-0} + 2^{-1} + 2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} + 2^{-2} + \cdots$$

□

Continuing from above, we have shown

$$\mathbb{E}[\ell(f(X))] \ge H(X) - \log \log 2|\mathcal{X}|.$$

Suppose we observe i.i.d. $X_1, \ldots, X_n$. The alphabet has size $|\mathcal{X}|^n$. Any description of these n outcomes requires at least $H(X^n) - \log(1 + n \log|X|)$ bits. Using independence, we have the following lower bound on the average number of bits per outcome.

$$\frac{1}{n}\mathbb{E}[\ell(f_n(X^n))] \ge H(X_1) - \frac{1}{n}\log(1 + n\log|\mathcal{X}|). = H(X) - O(\log(n)/n).$$

Suppose $\mathcal{X} = \{a, b, c\}$ and $f(a) = 0$, $f(b) = 1$, and $f(c) = 01$. Concatenating is no longer injective: 01101 could correspond to $abbc$ or cbc. If we want **uniquely decodable** f, then the upper bound of the previous proposition is 1.

Consider X being uniform on the above alphabet. Let $f(a) = \Lambda$, $f(b) = 0$, and $f(c) = 1$. Then $\mathbb{E}[\ell(f(X))] = 2/3 < 1.58 \approx H(X)$.

2 Asymptotic Equipartition Property

Suppose we observe i.i.d. $\mathrm{Ber}(1-p)$ random variables. The "typical" sequence has pn zeros and $(1-p)n$ ones. The probability of a *single* such sequence is $p^{pn}(1-p)^{(1-p)n} = 2^{-nH(X)}$.

Applying the law of large numbers to $f(x) = -\log p_X(x)$ gives

$$\frac{1}{n}\sum_i f(X_i) = \frac{1}{n}\log\frac{1}{p_{X^n}(x^n)} \xrightarrow{p} H(X)$$

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

Theorem 2.1.

$$-\frac{1}{n}\log p_{X^n}(x^n) \xrightarrow{p} H(X).$$

Last time we saw a few examples where the formula for entropy "magically" appeared. One was that if $f : \mathcal{X} \to \{0,1\}^*$ is injective, then $\mathbb{E}\ell(f(X)) \geq H(X) - \log\log 2|\mathcal{X}|$. Today we will show that there exists an f^* such that $\mathbb{E}\ell(f^*(X)) \lesssim H(X)$.

We also saw that if we observe an i.i.d. sequence of $\mathrm{Bern}(1-p)$ random variables, then the "typical" sequence of length n has $n(1-p)$ ones and np zeros, and moreover $p_{X^n}(x^n) \approx 2^{-nH(X)}$.

To formally define "typical," we will work backwards from the above example.

Definition 2.2. The **typical set** $A_\epsilon^{(n)} \subset \mathcal{X}^n$ is defined as

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

■

Proposition 2.3 (Properties of typical sets)**.**

1. $x \in A_\epsilon^{(n)} \iff H(X) - \epsilon \leq -\frac{1}{n}\log p_{X^n}(x^n) \leq H(X) + \epsilon$.
2. $\mathbb{P}(X^n \in A_\epsilon^{(n)}) \to 1$.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon')2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Proof. 1. follows by definition. 2. follows by the AEP/LLN.

3. follows by

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

4.

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

□

In summary, $A_\epsilon^{(n)}$ together has almost all the proability mass, the probability on sequences in $A_\epsilon^{(n)}$ is roughly uniform, and the cardinality is roughly $2^{nH(X)}$.

If $B_\delta^{(n)}$ is the smallest set with probability $\geq 1-\delta$, then $|B_\delta^{(n)}| \approx |A_\epsilon^{(n)}|$ in some sense. More precisely,

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

Proof.

$$\begin{aligned}
1 - \epsilon - \delta &\leq P(B_\delta^{(n)} \cap A_\epsilon^{(n)}) && \text{union bound} \\
&= \sum_{x^n \in A \cap B} p_{X^n}(x^n) \\
&\leq \sum_{x^n \in A \cap B} 2^{-n(H(X)-\epsilon)} \\
&\leq |B_\delta^{(n)}| 2^{-n(H(X)-\epsilon)} \\
|B_\delta^{(n)}| &\geq (1 - \epsilon - \delta)2^{n(H(X)-\epsilon)} \\
\frac{1}{n}\log|B_\delta^{(n)}| &> H(X) - \epsilon + \log(1 - \epsilon - \delta)
\end{aligned}$$

□

We now describe a scheme to describe X with $H(X)$ bits on average: the typical set encoding.

If we label all sequences in $A_\epsilon^{(n)}$, we need $\log|A_\epsilon^{(n)}|+1$ bits per label.

To encode $x \in A_\epsilon^{(n)}$, we put a flag 1 in front of the label (length $\log|A_\epsilon^{(n)}|+2$. If $x \notin A_\epsilon^{(n)}$, we put a flag 0 in front of the binary representation of the sequence (length $n(\log|X|+1)+1$).

$$\begin{aligned}\frac{1}{n}\mathbb{E}\ell(f(X^n)) &\le \frac{1}{n}P(A_\epsilon^{(n)})(\log|A_\epsilon^{(n)}|+2) + \frac{1}{n}(1-P(A_\epsilon^{(n)}))(n(\log|\mathcal{X}|+1)+1)\\ &\le (H(X)+\epsilon) + \frac{1}{n} + \delta(n)((\log|\mathcal{X}|+1)+1/n)\\ &\le H(X)+2\epsilon\end{aligned}$$

Although this scheme is not practical, we see that we get a matching upper bound to our earlier lower bound $H(X) - \log\log 2|\mathcal{X}|$ on the expected length.

3 Entropy rates of a stochastic process

For a random process $\{X_i\}$ the **entropy rate** is defined as

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

provided the limit exists.

Theorem 3.1 (Shannon-McMillan-Breiman)**.** For stationary ergodic processes,

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

with probability 1.

This is the AEP generalized to more general processes. AEP-like properties also generalize.

A **stationary** process has shift-invariant joint probabilities: $p_{X_1,\ldots,X_k} = p_{X_{\ell+1},\ldots,X_{\ell+k}}$. An **ergodic** process has time averages equalling ensemble averages in some sense, e.g. $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}X$ (a LLN-type property).

Non-ergodic process: choose a p-coin or a q-coin with equal probability ($p \neq q$) and flip it repeatedly. The space average is $(p+q)/2$, but the time average is either p or q.

Lemma 3.2. For a stationary process,

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

Proof. $0 \le H(X_{n+1} \mid X_1,\ldots,X_n) \le H(X_{n+1} \mid X_2,\ldots,X_n) = H(X_n \mid X_1,\ldots,X_{n-1})$. So this is a nonnegative decreasing sequence and therefore converges.

If $a_n \to a$ then the Cesaro means $b_n := \frac{1}{n}\sum_{i=1}^n a_i \to a$ also converge to the same limit. Applying this to $a_n := H(X_n \mid X_1,\ldots,X_{n-1})$ and $b_n = \frac{1}{n}\sum_{i=1}^n H(X_i \mid X_1,\ldots,X_{i-1}) = \frac{1}{n}H(X_1,\ldots,X_n)$ concludes. □

Markov chains are an example of stochastic processes. In particular if there is a stationary distribution π (satisfies $\pi P = \pi$), then the Markov chain is stationary and we may use the previous lemma. So $H(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_{n-1})$. We have

$$H(X_n \mid X_{n-1}) = \sum_i \pi(i)H(X_n \mid X_{n-1}=i) = \sum_i \pi(i)H(p_{i1},\ldots,p_{im})$$

The Second Law of Thermodynamics: the entropy of an isolated system is not decreasing. We might model such a system with a Markov process. [Example: Ehrenfest gas system?]

Suppose we start a Markov chain in two different states, and let their state distributions at time n be P_n and Q_n respectively, then

$$D(P_{n+1}\|Q_{n+1}) \le D(P_n\|Q_n).$$

If P_n is stationary, $D(P_n\|\pi)$ is decreasing.

4 Data compression

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

From AEP, it takes about $nH(X)$ bits to represent $X_1, \dots, X_n$. But this scheme (using typical set) is wildly impractical.

Let $c : \mathcal{X} \to \{0,1\}^*$ be a source code. We are interested in $L(c) = \sum_x p_X(x)\ell(c(x))$.

For example,

x	$p_X(x)$	$c(x)$	$\ell(c(x))$
a	$1/5$	00	2
b	$2/5$	01	2
c	$2/5$	1	1

has $L(c) = 1.6$ and $H(X) = 1.52$.

This defines the extension code c^* defined by $c^*(x_1 \cdots x_n) = c(x_1) \cdots c(x_n)$.

A **nonsingular** code c is an injective code. A code is **uniquely decodable** if c^* is nonsingular (injective).

A code is **prefix-free** (a.k.a. instantaneous) if no code word is a prefix of any other code word. Confusingly, such codes are sometimes called prefix codes.

$$\text{codes} \supset \text{nonsingular cords} \supset \text{uniquely decodable codes} \supset \text{prefix-free codes}.$$

Theorem 4.1 (Kraft inequality)**.** Any prefix code satisfies $\sum_{i=1}^N 2^{-\ell_i} \le 1$.

Also, for any numbers $(\ell_i)_{i=1}^N$ satisfying $\sum_{i=1}^N 2^{-\ell_i}$, there exists a prefix code with these codeword lengths.

Proof. Sort $\ell_1 \le \cdots \le \ell_N$. Draw a binary tree of code words. If $\ell_1 = 2$ for example, let the first code word be 00 and prune the children at that node. Then choose the next smallest available word of length ℓ_2 and so on. The inequality guarantees that this is possible and we don't run out of words.

We now show any prefix code satisfies the inequality. Consider inputting a sequence of random coin flips into a decoder which spits out x if the input is a code word.

$$1 \ge \mathbb{P}(\text{some } x \text{ comes out}) = \mathbb{P}\bigcup_{x\in\mathcal{X}}\{x \text{ comes out}\} = \sum_{x\in\mathcal{X}}\mathbb{P}(x \text{ comes out}) = \sum_{x\in\mathcal{X}} 2^{-\ell(c(x))}$$

□

Theorem 4.2 (McMillan inequality)**.** Any uniquely decodable code has codeword lengths satisfying $\sum_i 2^{-\ell_i} \le 1$.

We prove McMillan's inequality on the homework. Recall that for any f, we proved

$$\mathbb{E}[\ell(f(X))] = H(X) + D(p_X \| q) - \log \sum_x 2^{-\ell(f(x))}$$

where $q(x) \propto 2^{-\ell(f(x))}$. If c is uniquely decodable, then applying McMillan's inequality shows that $\mathbb{E}\ell(c(X)) \ge H(X)$.

How do we design good codes? Consider choosing ℓ_i to minimize $\sum_i p_X(i)\ell_i$ subject to $\sum_i 2^{-\ell_i} \le 1$. Then the optimal ℓ_is satisfy $\ell_i = -\log p_X(i)$. This is natural when considering $D(p_X\|q)$.

However, our code lengths must be integers, so we can consider instead $\ell_i = \lceil -\log p_X(i) \rceil$.

$$\sum_i p_X(i)\ell_i \le \sum_i p_X(i)(1 - \log p_X(i)) = H(X) + 1.$$

Theorem 4.3. Any uniquely decodable code c satisfies $\mathbb{E}\ell(c(X)) \ge H(X)$ (impossibility), and there exists a code c^* such that $\mathbb{E}\ell(c^*(X)) \le H(X) + 1$ (achievability).

The **Huffman code** is an "efficient" construction of the best prefix code.

Consider the following alphabet and distribution.

x	a	b	c
p_X	$1/5$	$2/5$	$2/5$

At each stage we group the two least probable symbols into one and add their probabilities. We start with $\{a\}$, $\{b\}$, and $\{c\}$. After the first stage we have $\{a, b\}$ w.p. $3/5$ and c w.p. $2/5$. Finally, we merge everything to get $\{a, b, c\}$. At each grouping, label the edges 0 and 1. Read backwards to get the code: $a = 00$, $b = 01$, and $c = 1$.

$\mathbb{E}\ell(c(X)) = 2/5 + 4/5 + 2/5 = 8/5 \approx 1.6$. While $H(X) \approx 0.722$.

Theorem 4.4. Huffman coding is optimal among all uniquely decodable codes. That is, if c^* is the Huffman code, then $\mathbb{E}\ell(c^*(X)) \leq \mathbb{E}\ell(c(X))$ for any uniquely decodable code c.

See book for proof.

Aside: consider the above distribution with code $a = 00$, $b = 1$, and $c = 0$. This is not uniquely decodable (although it is nonsingular). Its expected length is $6/5$, shorter than the Huffman code.

So we showed

$$H(X) \leq L^* \leq H(X) + 1.$$

One "trick" to get rid of the $+1$ is to group symbols together and give a code $\mathcal{X}^n \to \{0,1\}^*$. Then we have the following bounds on the number of bits to describe n symbols: $H(X^n) \leq \mathbb{E}\ell(c^*(X)) \leq H(X^n) + 1$. Dividing by n gives the following bounds on the number of bits per symbol:

$$H(X) \leq \frac{1}{n}\mathbb{E}\ell(c^*(X^n)) \leq H(X) + \frac{1}{n}.$$

The Huffman code for $\mathcal{X}^n$ requires building a tree with $|\mathcal{X}|^n$ leaves. We have a tradeoff between length of the code and computational complexity.

[See Lempel-Ziv for universal coding: variable length coding.]

This motivates arithmetic coding, which has complexity linear in n (instead of exponential as in the above discussion). We discuss **Shannon-Fano-Elias** coding. We will see that $\mathbb{E}\ell(c_{SFE}(X)) \leq H(X) + 2$.

Let the distribution of X be $(p_1, \ldots, p_3)$. We partition $[0, 1)$ into half-open intervals each of length p_i. [So, $[0, p_1)$, $[p_1, p_1 + p_2)$, and $[p_1 + p_2, p_1 + p_2 + p_3)$.] To encode i, take the midpoint of interval of length p_i, write it in binary, and truncate to $\lceil -\log p_i \rceil + 1$ bits. The truncation will always lie in the same interval, since truncation subtracts at most $2^{-(\lceil -\log p_i \rceil + 1)} \leq 2^{-(-\log p_i + 1)} = p_i/2$. We can also readily see $\mathbb{E}\ell(c_{SFE}(X)) \leq H(X) + 2$.

Suppose X takes values a, b, c with probabilities $0.6, 0.3, 0.1$ respectively. The intervals are $[0, 0.6), [0.6, 0.9), [0.9, 1)$. The mid points are $0.3 = 0.01\overline{0011}_2$, $0.75 = 0.11_2$, and $0.95 = 0.1111001\ldots$. So our code words are 01, 110, and 11110. It is clear this is not really optimal. However it is good for scaling up.

Consider another example where X takes values a, b, c with probabilities $1/2, 1/3, 1/6$ respectively. Then the product distribution is a distribution over 9 outcomes, e.g. ab has probability $1/6$.

Compare the interval representations for the $\mathcal{X}$ and $\mathcal{X}^2$ codes. In the latter, we would simply partition each of the three half-open intervals of the former into three to get a total of nine intervals.

In general if we want to encode a sequence of n bits, keep partitioning the intervals in the same way, and return any binary string in the small interval (e.g., truncating the midpoint to $\lceil -\log p_{X^n}(x^n) \rceil + 1$).

Note this procedure is linear in n.

Asymmetric Numeral System (ANS)? Coping with distributions that are less uniform; some probabilities very high. Skew binary representation? See paper.

5 Channel capacity

The communication problem:

$$\text{msg} \to \text{transmitter} \to \text{Channel (noisy)} \to \text{receiver} \to \text{msg}$$

Assumptions

1. Messages are random variables uniformly distributed over $\{1, \ldots, M\}$.
2. Channel has known statistical properties. Input X and output Y, know $P_{Y|X}$. Justification: in real life we can take measurements.

3. Discrete time, discrete alphabet, memoryless channel. Statistics of $P_{Y|X}$ do not change over time. [If X_1, X_2 are i.i.d. and input sequentially, then output Y_1, Y_2 are i.i.d.]

Example 5.1 (Binary symmetric channel)**.** The **binary symmetric channel** $\mathrm{BSC}(p)$ takes binary input $x \in \{0,1\}$ and outputs either x or $1-x$ with probability $1-p$ and p respectively. ■

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

$$W \to (M,n)\text{-block code} \xrightarrow{X^n(W)} P_{Y^n|X^n} \xrightarrow{Y^n} \text{decoder} \to \widehat{W}(Y^n)$$

The rate of an (M,n) code is $R = \frac{\log M}{n} = \frac{\text{number of information bits}}{\text{number of channel uses}} =$ bits/channel use. We also write an (M,n) code as a $(2^{nR}, n)$ code.

A good code should have high rate and good reliability, but these are in conflict.

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

$$\max_i \mathbb{P}(\widehat{W} \neq W_i \mid W = W_i) \to 0$$

We call this **reliable communication**.

Definition 5.2 (Operational definition of channel capacity)**.** The **capacity** C of a channel $P_{Y|X}$ is

$$\sup\{R : R \text{ is achievable}\}.$$

■

Consequently,

1. If $R < C$, then R is achievable. That is, for any $\epsilon > 0$, there exists an n and a $(2^{nR}, n)$ code with $\max_i \mathbb{P}(\widehat{W} \neq W_i \mid W = W_i) < \epsilon$.
2. If $R > C$, then R is not achievable. That is, there exists c such that for any sequence of $(2^{nR}, n)$ codes, $\liminf_{n\to\infty} \mathbb{P}(\widehat{W} \neq W_i \mid W = W_i) > c$.

Theorem 5.3 (Shannon's channel coding theorem)**.**

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

Properties of C.

1. $C \geq 0$.
2. $C \leq \min(\log|\mathcal{X}|, \log|\mathcal{Y}|)$. (Just note $I(X;Y) \leq H(X) \leq \log|\mathcal{X}|$.) This is equality when the channel is a deterministic map from $\mathcal{X} \to \mathcal{Y}$, and $|\mathcal{Y}| = |\mathcal{X}|$.
3. Recall $I(X;Y)$ is concave in P_X for fixed $P_{Y|X}$. Computing C is a convex optimization problem.

Example 5.4 (Binary symmetric channel)**.** Consider $\mathrm{BSC}(p)$. Note $I(X;Y) = H(Y) - H(Y \mid X) = H(Y) - H(p) \leq 1 - H(p)$ for any distribution on X. If $X \sim \mathrm{Ber}(1/2)$, then $Y \sim \mathrm{Ber}(1/2)$, which gives equality, so this is the maximizing distribution. $C = 1 - H(p)$. For example if $p = 0.1$, we have $C \approx 0.6$. ■

Example 5.5 (Binary erasure channel)**.** Consider the binary erasure channel $\mathrm{BEC}(p)$. Input $x \in \{0,1\}$ and output x with probability $1-p$ or e (erasure) with probability p. Letting E be the indicator for erasure, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) = H(Y) - H(p) \\ &= H(Y,E) - H(p) \\ &= H(E) + H(Y \mid E) - H(p) \\ &= H(Y \mid E) \\ &= pH(Y \mid E = 1) + (1-p)H(Y \mid E = 0) \\ &\leq (1-p). \end{aligned}$$

Taking $X \sim \mathrm{Ber}(1/2)$ gives equality again, so $C = 1 - p$. This makes sense: if we knew where the erasures were (proportion p of the time), we can send perfectly. What is surprising that we don't need to know where the erasures are. ■

Recall the setup.

$$\overset{W \in \{1,\dots,2^{nR}\}}{\longrightarrow} \text{encoder} \overset{X^n(W)}{\longrightarrow} \prod P_{Y_i|X_i} \overset{Y^n}{\longrightarrow} \text{decoder} \overset{\widehat{W}(Y^n)}{\longrightarrow}$$

A rate R is achievable if there exists a sequence of $(2^{nR}, n)$ codes such that $\max_i \mathbb{P}(\widehat{W} \neq W_i \mid W = W_i) \to 0$ as $n \to \infty$. The channel capacity C is a supremum of achievable rates; can be interpreted as the maximum rate at which information can be transmitted reliably.

We showed that the channel capacities for $\mathrm{BSC}(p)$ and $\mathrm{BEC}(p)$ are $1 - H(p)$ nad $1 - p$ respectively.

Heuristic explanation for channel capacity of $\mathrm{BEC}(p)$: First we argue an upper bound on achievable rates R. Suppose we have extra information: we know the location of the [on average] $\approx pn$ erasures. Then we could send [on average] $\approx (1-p)n$ bits reliably, i.e. $1 - p$ bits per channel use. This implies $R \leq 1 - p$ for achievable R.

Now, how do we actually achieve this without the actual information?

Consider a $G \in \{0,1\}^{n \times (1-p-\epsilon)n}$ that is full rank (over $\mathbb{F}_2$). Let $X^n(W) := GW \in \{0,1\}^n$ where $W \in \{0,1\}^{(1-p-\epsilon)n}$. Once we send it through the channel, [on average] proportion $\approx p$ bits are erased, so after the channel we have $n(1-p)$ bits that are not erased. Then we can invert the system since $1 - p - \epsilon < 1 - p$. Regarding finding a G, one can show that with i.i.d. Bernoulli entries, the resulting matrix is full rank with high probability. Note that in this scheme, the probability of error is the probability that the number of erasures is $> (p + \epsilon)n$.

Example 5.6 (Noisy typewriter)**.** $\mathcal{X} = \{A, \dots, F\}$, and for each letter X, the distribution $Y \mid X$ is equally likely to be X or $X + 1$.

$$I(X;Y) = H(Y) - H(Y \mid X) = H(Y) - 1 \leq \log(6) - 1 = \log 3.$$

If we choose P_X to be uniform on $\{A, C, E\}$, then we have equality $I(X;Y) = \log 3$. ■

Preview of Achievability in Channel Coding Theorem: all channels look like the noisy typewriter for n sufficiently large.

Example 5.7 (Additive Gaussian white noise)**.** Let $P_{Y|X} \sim \mathcal{N}(X, 1)$. Then the typical set is a ball centered at X^n on the order of $\sqrt{n}$. Then a simple encoding/decoding scheme is to choose (for the input distribution) a packing of the space so that these balls are disjoint; decoding just chooses the nearest center. ■

We now prove the theorem.

Proof of channel coding theorem (converse/impossibility). We begin with the converse: if R is achievable, then $R \leq C := \max_{P_X} I(X;Y)$. If R is achievable, there exists a sequence of $(2^{nR}, n)$ code with $\max_i \mathbb{P}(\widehat{W}^{(n)} \neq W_i \mid W^{(n)} = W_i) = \epsilon_n$ with $\epsilon_n \to 0$. This implies $\mathbb{P}(\widehat{W}^{(n)} \neq W^{(n)}) \leq \epsilon_n$ (average is less than maximum).

Note

$$\begin{aligned}
nR &= H(W^{(n)}) \\
&= H(W^{(n)} \mid \widehat{W}^{(n)}) + I(W^{(n)}; \widehat{W}^{(n)}) \\
&\leq 1 + \epsilon_n nR + I(W^{(n)}; \widehat{W}^{(n)}) && \text{Fano, } \widehat{W} \text{ is estimator of } W \\
&\leq 1 + \epsilon_n nR + I(X^n; Y^n) && \text{data proc. on } W^{(n)} \to X^n \to Y^n \to \widehat{W}^{(n)} \\
&= 1 + \epsilon_n nR + H(Y^n) - H(Y^n \mid X^n) \\
&\leq 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - H(Y^n \mid X^n) && \text{indep. bound (chain rule + cond. reduces entropy)} \\
&= 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X^n, Y^{i-1}) \\
&= 1 + \epsilon_n nR + \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X_i) && \text{memorylessness/Markov}
\end{aligned}$$

$$= 1 + \epsilon_n nR + \sum_{i=1}^{n} I(X_i; Y_i)$$
$$\leq 1 + \epsilon_n nR + nC.$$

Divide by n and take $n \to \infty$ (recall $\epsilon_n \to 0$) gives $R \leq C$.

Note that the above proof works fine even if we only have the weaker condition $\mathbb{P}(\widehat{W}^{(n)} \neq W^{(n)}) \leq \epsilon_n$ (average error over uniform on W). The other direction also only requires this weaker assumption.

Also note that the uniformity of W over the possible messages is crucial (appears in the step $nR = H(W^{(n)})$).

Note at the end of the above proof, we had $\epsilon_n \geq 1 - \frac{C}{R} - \frac{1}{nR} \approx 1 - \frac{C}{R}$. If $R > C$, then the probability of error is bounded from below. This is the weak converse.

There is a strong converse: if $R > C$, $\mathbb{P}(\widehat{W} \neq W) \geq 1 - 2^{-E(R,C)}$ where $E(R, C) > 0$ is some function(?)

Suppose R is very close to C. What do the above inequalities tell us?

- The use of Fano's inequality says we should use the best estimator $\widehat{W}$ of W...
- The use of data processing inequality implies $W \mapsto X^n$ and $Y^n \to \widehat{W}$ should be close to bijective.
- The next inequality implies the Y_i should be close to independent (i.e. the X_i should be close to independence).
- The last inequality implies that the distribution of X^n is close to i.i.d. from $\operatorname{argmax}_{P_X} I(X; Y)$. [Capacity-achieving input distribution (CAID).]

□

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

$$A_\epsilon^{(n)} = \left\{ (x^n, y^n) : \left| -\frac{1}{n} \log p_{X^n,Y^n}(x^n, y^n) - H(X,Y) \right| < \epsilon, \right.$$
$$\left| -\frac{1}{n} \log p_{X^n}(x^n) - H(X) \right| < \epsilon,$$
$$\left. \left| -\frac{1}{n} \log p_{Y^n}(y^n) - H(Y) \right| < \epsilon \right\}$$

If a rate R is close to C, then the function $X^n(W)$ ought to be "random" (specifically, i.i.d. from P_X^*, the optimizer of $I(X; Y)$).

Well-known codes like algebraic codes (Reed-Solomon, Gaulay, PCH) have a lot of structure and redundancy for the sake of simple decoding. However, this is at odds with the above intuition of a capacity-achieving code.

Today we will prove the achievability direction of the channel coding theorem.

Theorem 5.8 (Joint AEP)**.** Let $(X^n, Y^n) \sim P_{XY}^n$.

1. $\mathbb{P}((X^n, Y^n) \in A_\epsilon^{(n)}(X, Y)) \to 1$.
2. $|A_\epsilon^{(n)}(X, Y)| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. If $(\widetilde{X}^n, \widetilde{Y}^n) \sim (P_X \times P_Y)^n$, then
$$\mathbb{P}((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

The third statement is new. Independence causes the probability of being in the typical set to be vanishing. Note all pairings of typical X sequences and typical Y sequences are not necessarily jointly typical (in fact, most of them are not).

Proof. The first statement follows by the law of large numbers. The second statement follows by

$$1 \geq \sum_{x^n, y^n} p(x^n, y^n) \geq |A_\epsilon^{(n)}(X,Y)| 2^{-n(H(X,Y)+\epsilon)}.$$

For the third statement

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

□

Proof of achievability in channel coding theorem. We now prove that all rates $R < C$ are achievable. This is a non-constructive proof. We will use the probabilistic method: we will show that some object in a finite class $\mathcal{C}$ satisfies property $\mathcal{P}$ by exhibiting a distribution over $\mathcal{C}$, drawing an object from this distribution, and show that this object satisfies $\mathcal{P}$ with probability > 0.

In our setting $\mathcal{C}$ is a set of $(2^{nR}, R)$ codes and $\mathcal{P}$ is "code has small probability of error."

We will use random coding. Fix some P_X and $\epsilon > 0$ and R. Generate a $(2^{nR}, n)$ code at random according to P_X. The codebook is

$$\mathbb{C} = [X_i(w)]_{i,w} \in \mathbb{R}^{2^{nR} \times n}.$$

The wth row of $\mathbb{C}$ is the codeword for the wth message. Each letter of each code word is generated i.i.d. from P_X, so

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

We have decided the encoding scheme. For decoding, we will use typical set decoding. The receiver declares that $\widehat{W}$ was sent if both of the following happen.

1. $(X^n(\widehat{W}), Y^n) \in A_\epsilon^{(n)}(X,Y)$, that is, the received message Y^n is jointly typical with code word $X^n(\widehat{W})$.
2. No other index $k \neq \widehat{W}$ also satisfies $(X^n(k), Y^n) \in A_\epsilon^{(n)}(X,Y)$.

Otherwise, the decoder declares an error.

We now compute the expected probability of error for this scheme, averaged over codebooks $\mathbb{C}$ drawn from the above distribution.

The probability of error is

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

This probability is over the randomness of the channel, but the codebook fixed.

The average [over all messages] probability of error for a fixed code $\mathbb{C}$ is

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

[This is a function of $\mathbb{C}$, can be thought of as a conditional probability.]

The average over all codes is

$$\begin{aligned}
P_{error} &= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) P_\epsilon^{(n)}(\mathbb{C}) \\
&= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathbb{C}) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \lambda_w(\mathbb{C}) \\
&= \sum_{\mathbb{C}} \mathbb{P}(\mathbb{C}) \lambda_1(\mathbb{C}) && \text{rows of } \mathbb{C} \text{ are exchangeable} \\
&= \mathbb{P}(\text{error} \mid W = 1). && \text{error prob. averaged over all codes}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}(\text{error} \mid W = 1) &= \mathbb{P}(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}} \mid W = 1) \\
&\leq \mathbb{P}(E_1^c \mid W = 1) + \sum_{i=2}^{2^{nR}} \mathbb{P}(E_i \mid W = 1) \\
&= \mathbb{P}((X^n, Y^n) \notin A_\epsilon^{(n)}) + \sum_{i=2}^{2^{nR}} \mathbb{P}((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) \\
&\leq \epsilon + 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} && \text{for } n \text{ sufficiently large} \\
&\leq 2\epsilon && \text{for } n \text{ sufficiently large, if } R < I(X;Y) - 3\epsilon
\end{aligned}$$

For $i \neq 1$, Y^n is independent of $X^n(i)$ because Y^n came from $X_n(1)$.

So, the average [over codebooks and messages] probability of error is vanishing in n, so there exists some sequence of codebooks with vanishing probability of error [averaged over messages].

In conclusion, for any $R < C$ and $\epsilon > 0$ there exists a $(2^{nR}, n)$ code with average [over messages, not codebooks] probability of error $< \epsilon$.

"Almost every code is a good code, except the ones we construct." However, these codes are virtually impossible to decode. [Codebook is of exponential size, decoding needs to check all codewords.] □

Let X^n be i.i.d. $\mathrm{Bern}(1/2)$. Suppose Y^n is obtained by passing X^n through $\mathrm{BSC}(\alpha)$. [That is, Y_i is equal to X_i with probability $1 - \alpha$, and is flipped with probability α.] For any $b : \{0,1\}^n \to \{0,1\}$ (one-bit function), is it true that $I(b(X^n); Y^n) \leq 1 - H(\alpha) = I(X_1; Y_1)$? This is the "most informative function conjecture."

In the last few lectures we proved the channel coding theorem, which stated that the channel capacity is $\max_{P_X} I(X;Y)$.

- Converse: If R is achievable, then $R < C$. This holds using either definition of achievability (involving maximal probability of error $\max_i P(\widehat{W} \neq W_i \mid W = W_i)$ or average probability of error $P(\widehat{W} \neq W)$).
- Achievability: There exist $(2^{nR}, n)$ codes with $P(\widehat{W} \neq W) \to 0$ as $n \to \infty$ provided $R < C$.

We just need to tidy up the achievability result by proving the version with maximum probability of error. This is an application of Markov's inequality. Note

$$P(\widehat{W} \neq W) = \sum_{i=1}^{2^{nR}} P(\widehat{W} \neq W_i \mid W = W_i) P(W = W_i)$$

Since messages are uniformly distributed,

$$\#\{i : P(\widehat{W} \neq W_i \mid W = W_i) \geq \lambda\}/2^{nR} \leq \frac{P(\widehat{W} \neq W)}{\lambda}.$$

Taking $\lambda = 2P(\widehat{W} \neq W)$ gives

$$\#\{i : P(\widehat{W} \neq W_i \mid W = W_i) \geq 2P(\widehat{W} \neq W)\} \leq 2^{nR}/2.$$

If we throw away these bad messages (left-hand side), then the new rate is

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

In the new code, all the codewords have probability of error $\leq 2P(\widehat{W} \neq W)$ by definition, so

$$\max_i P(\widehat{W_0} \neq W_{0i} \mid W_0 = W_{0i}) \leq 2P(\widehat{W} \neq W) \to 0.$$

What is the rate of information sent from eye to brain? Measure signal X entering eye, signal Y entering brain, estimate $I(X;Y)$, gives upper bound on rate. [Estimating $I(X;Y)$ needs some shift to align due to delay, quantize time, etc.]

Suppose that we are the encoder that sends X_i through a channel, which sends Y_i to a decoder. What if we get **feedback**: we see Y_i (what the decoder receives)? We argued before that in the binary erasure channel, the capacity with feedback is the same. Is this true in general? [We know trivially that capacity with feedback must be at least as large as the capacity without feedback: just ignore the feedback.]

More explicitly, the encoder sends X_i, which can depend on W, as well as X^{i-1} and Y^{i-1}.

Theorem 5.9. The feedback does not improve the channel capacity. However, it can simplify the encoding scheme. It can also get us to capacity much more quickly. That is, $P_e(\text{best code}) \approx 2^{-nE}$ without feedback, but with feedback $P_e(\text{best code}) \approx 2^{-2^{nE'}}$.

Proof. We already know $C_{FB} \geq C$ so it suffices to prove the other direction.

$$nR = H(W) \qquad \text{messages are uniformly distributed}$$

$$\leq I(W;\widehat{W}) + nR\epsilon_n + 1 \qquad \text{Fano}$$

$$\leq I(W;Y^n) + nR\epsilon_n + 1 \qquad \text{data-processing, } W \to Y^n \to \widehat{W}$$

$$= H(Y^n) - H(Y^n \mid W) + nR\epsilon_n + 1$$

$$\leq \sum_i (H(Y_i) - H(Y_i \mid W, Y^{i-1})) + nR\epsilon_n + 1 \qquad \text{independence bound, chain rule}$$

$$= \sum_i (H(Y_i) - H(Y_i \mid W, Y^{i-1}, X_i)) + nR\epsilon_n + 1 \qquad X_i \text{ is function of } (W, Y^{i-1})$$

$$= \sum_i (H(Y_i) - H(Y_i \mid X_i)) + nR\epsilon_n + 1 \qquad Y_i \text{ is conditionally independent of all } \textit{past} \text{ things given } X_i$$

$$= \sum_i I(X_i;Y_i) + nR\epsilon_n + 1$$

$$\leq n \max_{P_X} I(X;Y) + nR\epsilon_n + 1.$$

Thus $R \le C$.

Remark: In the channel coding theorem, we had

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n \mid X^n) + n\epsilon_n \\ &\le \sum_i H(Y_i) - H(Y^n \mid X^n) + n\epsilon_n. \end{aligned}$$

We then used $H(Y^n \mid X^n) = \sum_i H(Y_i \mid X_i)$ because the Y_i were conditionally independent given the corresponding X_i. This is no longer the case with feedback. □

Preview of another problem (to be continued later):

Consider the following compression example. [No channel.] Suppose we have X^n and Y^n are correlated, and we want to encode each separately, and then a decoder takes both. (If the first encoder sents nR_X bits and the second sends nR_Y, then the rate is $R_X + R_Y$.) How much worse is this than encoding (X^n, Y^n) together? Chapter 15.4.

We say R_X and R_Y are **achievable** if there exists a sequence of functions $f : \mathcal{X}^n \to [2^{nR_X}]$ and $g : \mathcal{Y}^n \to [2^{nR_Y}]$ and $\phi : [2^{nR_X}] \times [2^{nR_Y}] \to \mathcal{X}^n \times \mathcal{Y}^n$ such that

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

as $n \to \infty$. $(\widehat{X}^n, \widehat{Y}^n) := \phi(f(X^n), g(Y^n))$. The **achievable rate region** is the closure of achievable rates.

Recall that if $R_X \ge H(X)$ then we can send the information losslessly. So the achievable rate region definitely contains $[R_X, \infty) \times [R_Y, \infty)$. Considering the special case where we can encode everything together, we see that the achievable rate region must lie in the region $\{R_X + R_Y \ge H(X, Y)\}$.

The answer: any rates satisfying $R_X \ge H(X \mid Y)$, $R_Y \ge H(Y \mid X)$, and $R_X + R_Y \ge H(X, Y)$.

In channel coding, high rate is desirable, but hard because of the channel. In compression, low rate is desirable (communicate using fewer bits) but hard. That is why achievable rates in channel coding has an upper bound, while in compression there is a lower bound.

We will discuss Problem 4 on the midterm.

(X^n, Y^n) are drawn i.i.d. from some distribution. Y^n is encoded into $f(Y^n) \in \{0,1\}^{nR}$. The decoder receives both $f(Y^n)$ and X^n, and gives an estimate $\widehat{Y}^n$.

The probability of error is $P_e^{(n)} = P(\widehat{y}^n(X^n, f(Y^n)) \ne Y^n)$. We say R is achievable if there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$.

First, we prove that if R is achievable, then $R \ge H(Y \mid X)$.

$$\begin{aligned} nR &\ge H(f(Y^n)) \\ &\ge H(f(Y^n) \mid X^n) && \text{conditioning reduces entropy} \\ &= H(Y^n, f(Y^n) \mid X^n) - H(Y^n \mid f(Y^n), X^n) && \text{chain rule} \\ &\ge H(Y^n, f(Y^n) \mid X^n) - (nP_e^{(n)} \log|\mathcal{Y}| + 1) \\ &= H(Y^n \mid X^n) + H(f(Y^n) \mid Y^n, X^n) - n\epsilon_n \\ &= nH(Y \mid X) + 0 - n\epsilon_n. \end{aligned}$$

Next, we prove a lemma about typical sets. We define $A_\epsilon^{(n)}(Y \mid x^n) = \{y^n : (x^n, y^n) \in A_\epsilon^{(n)}(X, Y)\}$. We prove $|A_\epsilon^{(n)}(Y \mid x^n)| \le 2^{nH(Y|X)+2\epsilon)}$.

$$\begin{aligned} 2^{-n(H(X)-\epsilon)} &\ge p(x^n) \\ &\ge \sum_{y^n : (x^n, y^n) \in A_\epsilon^{(n)}(X,Y)} p(x^n, y^n) \\ &\ge \sum_{y^n : (x^n, y^n) \in A_\epsilon^{(n)}(X,Y)} 2^{-n(H(X,Y)+\epsilon)} \\ &= |A_\epsilon^{(n)}(Y \mid x^n)| 2^{-n(H(X,Y)+\epsilon)} \end{aligned}$$

6 Differential entropy

The **differentiable entropy** of a continuous random variable with density f is $h(X) := -\int f(x)\log f(x)\,dx = -\mathbb{E}\log f(X)$.

If X is discrete and Y is continuous, then $I(X;Y) = H(X) - H(X \mid Y) = h(Y) - h(Y \mid X)$.

Example 6.1 (Uniform)**.** If X is uniform on $[0,a]$, then $h(X) = -\int_0^a \frac{1}{a}\log\frac{1}{a}\,dx = \log a$. In particular, $h(X)$ can be negative! ■

Why can entropy be negative? Consider approximating the integral by $-\sum_{i=-\infty}^{\infty}\frac{1}{N}f(i/N)\log f(i/N)$. Consider $[X]_N$, a discretized version of f taking values $1/N$ with probabilities $f(i/N)/N$. Then the above approximation to the integral is $H([X]_N) - \log N = H([X]_N) - H([U]_N)$, where $[U]_N$ is the discretization of a $\mathrm{Unif}(0,1)$ random variable. This is the "differential" in the name: it is in some sense the difference in discrete entropy of quantized versions of X and U.

The relative entropy is $D(P\|Q) = \int p(x)\log\frac{dp}{dq}(x)\,dx = \mathbb{E}_P\log\frac{dp}{dq}(X)$. Note that we can rewrite this as $\int \frac{dp}{dq}(x)\log\frac{dp}{dq}(x)\,dq(x)$. So, differential entropy can be written as $-D(P\|\,dx)$ where "dx" denotes the Lebesgue measure.

Example 6.2 (Gaussian)**.** Let $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/2\sigma^2}$.

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

■

Joint density is $h(X_1,\dots,X_n) = -\int f\log f$ where f is the joint density.

Example 6.3 (Multivariate Gaussian)**.** Let $f(x) = (2\pi)^{-n/2}|K|^{-1/2}\exp\big(-\frac{1}{2}(x-\mu)^\top K^{-1}(x-\mu)\big)$.

$$\begin{aligned} h(X) &= \int f\log\frac{1}{f} \\ &= \frac{1}{2}\log(2\pi)^n|K| + \frac{1}{2}\log(e) + \frac{1}{2}\log(e)\mathbb{E}(X-\mu)K^{-1}(X-\mu) \\ &= \frac{1}{2}\log(2\pi)^n|K| + \frac{n}{2}\log e \qquad \text{trace trick} \\ &= \frac{1}{2}\log(2\pi e)^n|K|. \end{aligned}$$

Note that μ does not appear. This is because entropy is invariant to shifting. ■

Conditional differential entropy is $h(Y \mid X) = -\int f_{X,Y}(x,y)\log f_{Y|X}(y \mid x)\,dy\,dx = -\mathbb{E}\log f_{Y|X}(Y \mid X)$.

Chain rule: $h(Y,X) = h(X) + h(Y \mid X)$.

Relative entropy: if $\mathrm{supp}(f) \subset \mathrm{supp}(g)$, $D(f\|g) = \int f\log\frac{f}{g} \geq 0$ by Jensen's inequality.

Mutual information is $I(X;Y) = D(f_{XY}\|f_Xf_Y) = h(X) - h(X \mid Y) \geq 0$. From this we see conditioning still decreases entropy.

We still have $h(X_1,\dots,X_n) = \sum_{i=1}^n h(X_i \mid X_1,\dots,X_{i-1}) \leq \sum_{i=1}^n h(X_i)$.

Also, $h(X+c) = h(X)$ and $h(aX) = h(X) + \log|a|$. [Recall if $Y = aX$ then $f_Y(y) = f_X(y/a)/|a|$.] More generally, for a matrix A, $h(AX) = h(X) + \log|A|$.

Aside:

Proposition 6.4 (Entropy power inequality)**.** Let X and Y be independent random vectors on $\mathbb{R}^n$.

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

Suppose X is uniform on some set A, so $f_X(x) = 1/\operatorname{vol}(A)\mathbf{1}_A$. Then $h(X) = \log \operatorname{vol}(A)$. Similarly if Y is uniform on another set B, then $h(Y) = \log \operatorname{vol}(B)$. Then, the entropy power inequality implies

$$2^{2h(X+Y)/n} \geq \operatorname{vol}(A)^{2/n} + \operatorname{vol}(B)^{2/n}.$$

The Brunn-Minkowski inequality states $\operatorname{vol}(A+B)^{1/n} \geq \operatorname{vol}(A)^{1/n} + \operatorname{vol}(B)^{1/n}$, where $A + B$ is the Minkowski sum $\{a + b : a \in A, b \in B\}$. If we take $2^{2h(X+Y)/n} \approx \operatorname{vol}(A+B)^{2/n}$ (note this is not true, due to the convoluting), then we see that the entropy power inequality suggests a stronger inequality than the Brunn-Minkowski inequality.

Rough volume argument for strong converse of channel coding: if $R > C$, The number of typical y^n is $\approx 2^{nH(Y)}$. For each x^n, number of conditionally typical y^n is $2^{nH(Y|X)}$. If these sets are disjoint, the number of sets in y^n is about $2^{nI(X;Y)}$. If we have $R > C$ then we have overlap.

If X is discrete and Y is continuous, no notion of joint entropy. However, we can talk about mutual information.

$$I(X;Y) = H(X) - H(X \mid Y) = h(Y) - h(Y \mid X).$$

For general random variables, we have another equivalent definition of mutual information.

$$I(X;Y) = \sup_P I([X]_P; [Y]_P),$$

where the supremum is over all partitions. Note that this is essentially the definition of Lebesgue integration from approximation by simple functions. For continuous Y, we recover the earlier definition.

$$I(X;Y) = \sup_N H([Y]_N) - H([Y]_N \mid [X]_N) \approx h(Y) + H([U]_N) - h(Y \mid [X]_N) - H([U]_N) \to h(Y) - h(Y \mid X).$$

The fact that mutual information can be defined between discrete and continuous random variables is good in practice. Consider a codeword $X^n \in [2^{nR}]$ being sent through a channel; such channels usually produce continuous output.

Last time we showed for $X \sim N(\mu, K)$, we have $h(X) = \frac{1}{2}\log(2\pi e)^n |K|$.

Theorem 6.5. For any random variable Y with covariance K,

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

Moreover, equality holds if and only if Y is Gaussian.

So, the Gaussian distribution maximizes entropy under a second moment constraint.

Aside: Note that for the discrete case, uniform distribution over finite alphabet maximizes entropy without moment conditions. It does not usually make sense to impose moment conditions since we usually do not care about the values of X, unlike the continuous case. For nonnegative integers with a mean constraint, geometric distribution maximizes entropy.

Proof. Let $\phi(x) = (2\pi)^{-n/2}|K|^{-1/2}e^{-x^\top K^{-1}x/2}$ be the Gaussian density and f arbitrary. [WLOG both f and ϕ have zero mean.] Then $-\log_e \phi = \frac{1}{2}\log[(2\pi)^n|K|] + \frac{1}{2}x^\top K^{-1}x$.

$$
\begin{aligned}
0 &\le D_e(f\|\phi) \\
&= \int f \log_e \frac{f}{\phi} \\
&= -\int f \log_e \phi - h_e(f) \\
&= \mathbb{E}_{X\sim f}[-\log_e \phi(X)] - h_e(f) \\
&= \frac{1}{2}\log[(2\pi)^n|K|] + \frac{1}{2}\underbrace{\mathbb{E}_{X\sim f} X^\top K^{-1} X}_{=n} - h_e(f) \\
&= \frac{1}{2}\log[(2\pi)^n|K|] + \frac{1}{2}\log_e e^n - h_e(f) \\
&= \frac{1}{2}\log_e[(2\pi e)^n|K|] - h_e(f) \\
&= h_e(\phi) - h_e(f).
\end{aligned}
$$

□

7 Gaussian channel

The Gaussian channel takes input X and outputs $Y = X + Z$ where $Z \sim N(0, \sigma^2)$ is independent of X.

We have $\sup_{P_X} I(X;Y) = \infty$ because we do not have constraints on X. We could spread the distribution of X so widely such that the Gaussian Z does not perturb by much, and Y can easily be decoded.

We consider instead $\sup_{P_X:\mathbb{E}X^2\le P} I(X;Y)$. [WLOG we assume X is zero mean; shifting does not change anything.] We have

$$
\begin{aligned}
I(X;Y) = h(Y) - h(Y \mid X) &= h(Y) - h(Z) \\
&= h(Y) - \frac{1}{2}\log 2\pi e\sigma^2 \\
&\le \frac{1}{2}\log 2\pi e(\sigma^2 + P) - \frac{1}{2}\log 2\pi e\sigma^2 \\
&= \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right).
\end{aligned}
$$

Equality is attained by $X \sim N(0, P)$, so $C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$. The ratio p/σ^2 is the signal-to-noise (SNR).

For the Gaussian channel, a $(2^{nR}, n, P)$ code c is a map from $w \in [2^{nR}]$ and outputs $X^n(w) \in \mathbb{R}^n$ such that $\|X^n(w)\|^2 \le nP$. [The idea is that we have limited energy; we cannot amplify arbitrarily large.]

$$
\begin{aligned}
nR &\le H(W) \\
&\le I(X^n; Y^n) + nP_e^{(n)}R + 1 && \text{data proc., Fano} \\
&\le \sum_{i=1}^n I(X_i;Y_i) + 1 + nP_e^{(n)}R \\
R &\le \frac{1}{n}\sum_{i=1}^n I(X_i;Y_i) + \epsilon_n
\end{aligned}
$$

We cannot take the supremum over P_X because we have the constraint $\|X^n(w)\|^2 \le nP$. We use the fact that for fixed $P_{Y|X}$, the map $P_X \mapsto I(X;Y)$ is concave. We define $P_X = \frac{1}{n}\sum_{i=1}^n P_{X_i}$. Then $P_Y = \int P_{Y|X}\, dP_X$. Then

$$
\frac{1}{n}\sum_{i=1}^n I(X_i;Y_i) \le I(X;Y) \le \max_{P_X:\mathbb{E}X^2\le P} I(X;Y)
$$

because $\mathbb{E}X^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^2 \le P$. With a little more work, the constraint $\|X^n(W)\|^2 \le nP$ could be relaxed to hold on average over messages W.

Let $X_1, X_2, \ldots$ be i.i.d. with density f, mean zero, and second moment equal to 1. The central limit theorem states

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} N(0,1).$$

Note that S_n all have the same second moment 1. We suspect $h(S_n)$ is increasing, since the limiting distribution is Gaussian.

Recall the entropy power inequality, which implies

$$2^{2h((X_1+X_2)/\sqrt{2})} = 2^{2(h(X_1+X_2)-\log 2)} \ge \frac{1}{2}(2^{2h(X_1)} + 2^{2h(X_2)}) = 2^{2h(X_1)}.$$

This argument shows $h(S_{2^k}) \ge h(S_{2^\ell})$ for $k \ge \ell$.

But does $h(S_n)$ increase monotonically? This was an open problem since Shannon, and solved in 2004.

Note that $h(S_m) \ge h(S_n)$ is equivalent to $D(S_m \| N(0,1)) \ge D(S_n \| N(0,1))$. [Convergence in entropy implies convergence in distribution. See Pinsker?] So a strong central limit theorem holds.

8 Entropy power inequality

Theorem 8.1.

$$2^{2h(X+Y)} \ge 2^{2h(X)} + 2^{2h(Y)}.$$

- 1948: proposed by Shannon (proof was wrong)
- 1959: Stam (semigroup / Fisher information)
- 1965: Blachman (same technique)
- 1991: Carlen, Soffer (same technique)
- ?: Dembo, Cover

Let $U = \sqrt{\lambda}X$ and $V = \sqrt{1-\lambda}Y$.

$$\begin{aligned}
2^{h(U+V)} &\ge 2^{2h(U)} + 2^{2h(V)} \\
2^{2h(\sqrt{\lambda}X+\sqrt{1-\lambda}Y)} &\ge \lambda 2^{2h(X)} + (1-\lambda)2^{2h(Y)} \qquad h(\sqrt{\lambda}X) = h(X) + \frac{1}{2}\log\lambda \\
&\ge 2^{2\lambda h(X)+2(1-\lambda)h(Y)} \qquad \text{Jensen}
\end{aligned}$$

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

This last inequality is **Lieb's inequality**, and we have shown it is a consequence of the entropy power inequality. This is actually equivalent to the entropy power inequality: choose $\lambda = \frac{2^{2h(U)}}{2^{2h(U)}+2^{2h(V)}}$.[1] The latter form is more convenient for proving, but the original form is more convenient for applications.

Without loss of generality, we may assume the densities of X and Y do not vanish. [Else convolve with a Gaussian with low variance; does not change much.]

[1] Suppose we want to prove the n-dimensional EPI $2^{2h(U+V)/n} \ge 2^{2h(U)/n} + 2^{2h(V)/n}$. Let $\lambda = \frac{2^{2h(U)/n}}{2^{2h(U)/n}+2^{2h(V)/n}}$.

$$\begin{aligned}
h(U+V) &= h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \\
&\ge \lambda h(X) + (1-\lambda)h(Y) \\
&= \lambda\Big(h(U) - \frac{n}{2}\log\lambda\Big) + (1-\lambda)\Big(h(Y) - \frac{n}{2}\log\lambda\Big) \\
&= \frac{n}{2}\log\Big(2^{2h(U)/n} + 2^{2h(V)/n}\Big).
\end{aligned}$$

We will be using results from optimal transport theory (but we will derive things from scratch). This is an adaptation of Olivier Rioul's proof (2016).

Let F_X and F_Y be the cdfs of X and Y respectively. If $X^* \sim N(0,1)$, then $\Phi(x^*)$ is uniform on $[0,1]$, and $F_X^{-1}(\Phi(X^*))$ is distributed as X. So let $T_X = F_X^{-1} \circ \Phi$ so that $T_X(X^*)$ is distributed as X. Note that T_X is an increasing function. [We assumed the densities are nonvanishing so the CDFs are strictly increasing.] Thus $T_X' > 0$. Define T_Y similarly.

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

Let $\widetilde{X}, \widetilde{Y} \sim N(0,1)$ be i.i.d. Then we can write

$$X^* = \sqrt{\lambda}\widetilde{X} - \sqrt{1-\lambda}\widetilde{Y}$$
$$Y^* = \sqrt{1-\lambda}\widetilde{X} + \sqrt{\lambda}\widetilde{Y}$$

[This is a unitary transformation; Gaussian distribution is rotation invariant.]

Consider

$$\Theta_{\widetilde{y}}(\widetilde{x}) = \sqrt{\lambda}T_X(x^*) + \sqrt{1-\lambda}T_Y(y^*).$$

Then

$$\Theta_{\widetilde{Y}}(\widetilde{X}) \stackrel{d}{=} \sqrt{\lambda}X + \sqrt{1-\lambda}Y.$$

We also have

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

Let f be the density of $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$. By the change of variables formula,

$$f_{\widetilde{y}}(\widetilde{x}) = f(\Theta_{\widetilde{y}}(\widetilde{x}))\Theta_{\widetilde{y}}'(\widetilde{x})$$

is a density.

$$\begin{aligned}
h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= \mathbb{E}\log\frac{1}{f(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)} \\
&= \mathbb{E}\log\frac{1}{f(\Theta_{\widetilde{Y}}(\widetilde{X}))} \\
&= \mathbb{E}\log\frac{\Theta_{\widetilde{y}}'(\widetilde{x})}{f_{\widetilde{Y}}(\widetilde{x})} \\
&= h(\widetilde{X}) + \mathbb{E}\log\frac{\phi(\widetilde{X})}{f_{\widetilde{Y}}(\widetilde{X})} + \mathbb{E}\log\Theta_{\widetilde{Y}}'(\widetilde{X}) \\
&= h(\widetilde{X}) + \mathbb{E}_{\widetilde{Y}}\mathbb{E}_{\widetilde{X}}\left[\log\frac{\phi(\widetilde{X})}{f_{\widetilde{Y}}(\widetilde{X})} \mid \widetilde{Y}\right] + \mathbb{E}\log\Theta_{\widetilde{Y}}'(\widetilde{X}) \\
&= h(\widetilde{X}) + \mathbb{E}_{\widetilde{Y}}D(g\|f_{\widetilde{Y}}) + \mathbb{E}\log\Theta_{\widetilde{Y}}'(\widetilde{X}) \\
&\geq h(\widetilde{X}) + \mathbb{E}\log\Theta_{\widetilde{Y}}'(\widetilde{X}) \\
&\geq h(\widetilde{X}) + \lambda\mathbb{E}\log T_X'(\widetilde{X}) + (1-\lambda)\mathbb{E}\log T_Y'(\widetilde{Y}) \qquad \text{concavity of logarithm} \\
&= \lambda(h(\widetilde{X}) + \mathbb{E}\log T_X'(\widetilde{X})) + (1-\lambda)(h(\widetilde{Y}) + \mathbb{E}\log T_Y'(\widetilde{Y})) \\
&= \lambda h(X) + (1-\lambda)h(Y).
\end{aligned}$$

The last step is due to the change of variables $\phi(\widetilde{x}) = f_X(T_X(\widetilde{x}))T_X'(\widetilde{X})$,

$$h(\widetilde{X}) + \mathbb{E}\log T_X'(\widetilde{X}) = \mathbb{E}\log\frac{T_X'(\widetilde{X})}{\phi(\widetilde{X})} = \mathbb{E}\log\frac{1}{f_X(T_X(\widetilde{x}))} = \mathbb{E}\log\frac{1}{f_X(X)} = h(X).$$

To relax the condition that the densities of X and Y are non-vanishing, we use an approximation $h(X + \sqrt{\epsilon}Z) \to h(X)$ as $\epsilon \to 0$.

We proved the entropy power inequality in dimension 1. In higher dimensions it takes a similar form in that the constants do not degrade; it is *dimension free*.

Theorem 8.2 (Conditional EPI)**.** If X and Y are conditionally independent given U.

$$2^{2h(X+Y|U)} \geq 2^{2h(X|U)} + 2^{2h(Y|U)}.$$

Proof. By the EPI,

$$\begin{aligned} 2h(X+Y \mid U=u) &\geq \log\left(2^{2h(X|U=u)} + 2^{2h(Y|U=u)}\right) \\ 2h(X+Y \mid U) &\geq \log\left(2^{2h(X|U)} + 2^{2h(Y|U)}\right) && \text{concavity of log-sum-exp.} \end{aligned}$$

□

Theorem 8.3 (EPI in n dimensions)**.** If X^n and Y^n are random vectors in $\mathbb{R}^n$,

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

Proof.

$$\begin{aligned} h(X^n + Y^n) &= h(X_n + Y_n) + h(X^{n-1} + Y^{n-1} \mid X_n + Y_n) \\ 2h(X^{n-1} + Y^{n-1} \mid X_n, Y_n) &\geq (n-1)\log\left(2^{\frac{2}{n-1}h(X^{n-1}|X_n,Y_n)} + 2^{\frac{2}{n-1}h(Y^{n-1}|X_n,Y_n)}\right) \\ 2h(X^{n-1} + Y^{n-1} \mid X_n + Y_n) &\geq (n-1)\log\left(2^{\frac{2}{n-1}h(X^{n-1}|X_n)} + 2^{\frac{2}{n-1}h(Y^{n-1}|Y_n)}\right) \\ 2h(X_n + Y_n) &\geq \log\left(2^{2h(X_n)} + 2^{2h(Y_n)}\right) \\ \frac{2}{n}h(X^n + Y^n) &\geq \log\left(2^{\frac{2}{n}h(X^n)} + 2^{\frac{2}{n}h(Y^n)}\right) && \text{concavity of log-sum-exp} \end{aligned}$$

□

Consider a adversarial channel who sees the distribution of X, chooses distribution Z and outputs $Y = X + Z$. We want to find the capacity

$$\sup_{P_X} \inf_{P_Z} I(X; X+Z)$$

subject to $\mathbb{E}X^2 \leq \sigma_X^2$ and $\mathbb{E}Z^2 \leq \sigma_Z^2$.

$$\begin{aligned} I(X; X+Z) &= h(X+Z) - h(Z) \\ &\geq h(X+Z) - h(Z^*) && Z^* \sim N(0, \sigma_Z^2) \\ &\geq h(X^* + Z^*) - h(Z^*) && \text{EPI} \\ &= \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) \end{aligned}$$

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

Thus, we have the Gaussian channel saddle point property:

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

In other words,

$$I(X;X+Z^*) \le I(X;X+Z) \le I(X^*;X^*+Z).$$

So non-Gaussian channels have higher capacity.

Getting Gaussian codes to work for any channel: perform unitary transformation before sending to channel, and then perform inverse on output. CLT?

---

9 Rate distortion theory

Lossy compression. Encoder observes X^n, sends nR bits to decoder, who then outputs estimate $\widehat{X}^n$.

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

If $R < H(X)$, then what can we say? Trade-off between dimension reduction (rate, number of bits in the representation) and fidelity of the reconstruction.

To measure fidelity, we need some **distortion function** (measure) $d : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}^+$. For convenience we sometimes consider bounded distortion functions that satisfy $d_{\max} = \max_{x,\widehat{x}} d(x,\widehat{x}) < \infty$. For example, $(x-\widehat{x})^2$ is unbounded. Most results generalize to the unbounded case.

Some distortion functions are

- **Hamming distortion** $d(x,\widehat{x}) = \mathbf{1}[x \neq \widehat{x}]$, used in the case where $\mathcal{X} = \widehat{\mathcal{X}}$ (or some subset relationship). Note $\mathbb{E}d(X,\widehat{X}) = P(X \neq \widehat{X})$.
- **Squared error / quadratic loss** $d(x,\widehat{x}) = (x-\widehat{x})^2$. Again, this is unbounded on $\mathbb{R}^2$.

We can extend distortion functions to sequences by $d(x^n,\widehat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i,\widehat{x}_i)$, the average per-symbol distortion. Other possibilities that we will not consider include $d(x^n,\widehat{x}^n) = \max_i d(x_i,\widehat{x}_i)$.

We have a encoding function $f_n : \mathcal{X}^n \to [2^{nR}]$ and a decoding function $g_n : [2^{nR}] \to \widehat{\mathcal{X}}^n$.

The central quantity is the **expected distortion**

$$\mathbb{E}d(X^n, \underbrace{g_n(f_n(X^n))}_{\widehat{X}^n}) = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))).$$

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

$$\lim_{n\to\infty} \mathbb{E}d(X^n, g_n(f_n(X^n))) \le D.$$

Note that if (R,D) is achievable, then (R,D') and (R',D) are also achievable if $D' \ge D$ and $R' \ge R$.

The achievable rate distortion region is also convex: if you have two codes, you can flip a coin to choose which code to use for a particular block. The rate and distortion will just be the convex combinations.

The point where the boundary of the achievable region hits the $D = 0$ axis is $(H(X), 0)$. Where the boundary hits $R = 0$ is $(0, \min_{\widehat{x}} \mathbb{E}d(X,\widehat{x}))$ (output the least offending guess on average). If $d_{\max}$ exists, it is larger than this minimum.

Note that the achievability definition can be restated $\epsilon > 0$ such that $P(d(X^n, g_n(f_n(X^n)) > D+\epsilon) \to 0$ for all $\epsilon > 0$. (???)

The **rate distortion function** is $R(D) = \inf\{R : (R,D) \text{ achievable}\}$, the lower boundary of the achievable rate distortion region.

Theorem 9.1. For $X_i \sim P_X$ i.i.d., $|\mathcal{X}|, |\widehat{\mathcal{X}}| < \infty$, and distortion bounded by $d_{\max}$,

$$R(D) = \min_{P_{\widehat{X}|X} : \mathbb{E}d(X,\widehat{X}) \le D} I(X;\widehat{X}).$$

In channel coding, the channel is fixed and we control the distribution over the input. In this setting, the input P_X is given, and we control the encoding/decoding scheme that gives the output.

Recall $P_{\widehat{X}|X} \mapsto I(X;\widehat{X})$ is convex for fixed P_X, so this is a convex optimization problem. [Note that the constraint on the expected distortion is a linear constraint: $\mathbb{E}d(X,\widehat{X}) = \sum_{x,\widehat{x}} p(x)p(\widehat{x} \mid x)d(x,\widehat{x})$.]

In channel coding, if we did not choose too many inputs, their typical images under the channel would hopefully be disjoint. This is a packing argument. We could use a volume argument to estimate how many inputs we can send.

In our setting, we have some subset of $\widehat{\mathcal{X}}^n$ of size 2^{nR}, and we consider the "preimage" of inputs in $\mathcal{X}^n$ that are within distortion D of these outputs $\widehat{X}(i)$. We want enough in the output space so that their "preimages" cover $\mathcal{X}^n$. This is in some sense a dual of channel coding. Note that R controls the number of "preimages," and with lower R, it becomes harder to cover. D controls the size of each "preimage" and lower D makes it harder to cover.

Example 9.2. Let $X \sim \mathrm{Ber}(p)$ with $p \le 1/2$. Let $\oplus$ be the XOR operation, and consider the Hamming distortion. Lower bounding $I(X,\widehat{X})$ subject to $\mathbb{E}d(X,\widehat{X}) \le D$ gives

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

$H(p) - H(D)$ is the mutual information corresponding to the BSC with transition probability D taking some input $\widetilde{X}$ and having output distribution Y following $(p, 1-p)$. Indeed, $I(X;Y) = H(Y) - H(Y \mid X) = H(p) - H(D)$. How do we use this information to achieve this rate in our original problem?

Put $\widehat{X}$ through a $\mathrm{BSC}(D)$ channel so that X has distribution $(p, 1-p)$. Using Bayes's Rule gives $P(\widehat{x} = 0) = \frac{1-p-D}{1-2D}$.

So,

$$R(D) = \begin{cases} H(p) - H(D) & D \le p \\ 0 & D > p \end{cases}$$

If $D \ge p$, we can simply output 0 all the time, and then the expected distortion is p. ■

Example 9.3. Consider a Gaussian random variable $X \sim N(0,\sigma^2)$. How would we do a one-bit quantization? Let $f(X) := \widehat{x}\mathbf{1}[X \ge 0] - \widehat{x}\mathbf{1}[X < 0]$. If $\widehat{x} = \sqrt{\frac{2}{\pi}}\sigma$, then $\mathbb{E}(X - f(X))^2 = \frac{\pi-2}{\pi}\sigma^2 \approx 0.36\sigma^2$.

We want to find $R(D) = \min_{P_{\widehat{X}|X}:\mathbb{E}(X-\widehat{X})^2 \le D} I(X;\widehat{X})$.

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

We have shown a lower bound. Is equality attained?

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

Consider a Gaussian channel $X = \widehat{X} + Z$ where $\widehat{X} \sim N(0,\sigma^2 - D)$ and $Z \sim N(0,D)$, then $X \sim N(0,\sigma^2)$. Then $I(X;\widehat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}$. So, $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$.

Now compare with our one-bit quantizer. With the same second fidelity constraint $D = 0.36\sigma^2$, the optimal rate is $R(0.36\sigma^2) = \frac{1}{2}\log\frac{1}{0.36} \approx 0.737$. This is a significant improvement over the rate 1 of the one-bit quantizer. It is suboptimal to quantize on a symbol-by-symbol basis; we gain by quantizing on blocks. ■

Suppose $X \sim N(0,1)$ is the input to a neural network which outputs $\widehat{X}$. Suppose we have $\mathbb{E}(X-\widehat{X})^2 = 1/2$. The information flow through any layer is $\geq R(1/2) = 1/2$.

We now prove the theorem.

Proof of achievability. Fix $P_{\widehat{X}|X}$ such that $\mathbb{E}d(X,\widehat{X}) \leq D$. We want to show that there exists a sequence of $(2^{nR}, n)$ codes that have rate $R \approx I(X;\widehat{X})$ and achieve $\mathbb{E}[d(X^n, g_n(f_n(X^n))) \to D$ as $n \to \infty$.

We define a distortion-typical set.

$$\begin{aligned}
A_{d,\epsilon}^{(n)} := \Big\{(x^n,\widehat{x}^n) : &\left|-\frac{1}{n}\log p_(x^n,\widehat{x}^n) - H(X,\widehat{X})\right| < \epsilon,\\
&\left|-\frac{1}{n}\log p(x^n) - h(X)\right| < \epsilon,\\
&\left|-\frac{1}{n}\log p(\widehat{x}^n) - h(\widehat{X})\right| < \epsilon,\\
&|d(x^n,\widehat{x}^n) - \mathbb{E}d(X,\widehat{X})| < \epsilon\Big\}\\
= A_\epsilon^{(n)}(X,\widehat{X}) &\cap \{(x^n,\widehat{x}^n) : |d(x^n,\widehat{x}^n) - \mathbb{E}d(X,\widehat{X})| < \epsilon\}\\
\subseteq A_\epsilon^{(n)}(X,&\widehat{X}).
\end{aligned}$$

1. We have $P(A_{d,n}^{(\epsilon)}) \to 1$ since the two sets in the intersection have probability tending to 1, both by the weak law of large numbers.
2. $p(\widehat{x}^n) \geq p(\widehat{x}^n \mid x^n) 2^{-n(I(X;\widehat{X})+3\epsilon)}$ for all $(x^n,\widehat{x}^n) \in A_{d,\epsilon}^{(n)}$.

$$p(\widehat{x}^n \mid x^n) = p(\widehat{x}^n)\frac{p(\widehat{x}^n,x^n)}{p(x^n)p(\widehat{x}^n)} \geq p(\widehat{x}^n)2^{-n(H(X,\widehat{X})-H(X)-H(\widehat{X})+3\epsilon)} = p(\widehat{x}^n)2^{n(I(X;\widehat{X})+3\epsilon)}$$

3. If $0 \leq x,y \leq 1$ and $n \geq 0$, then $(1-xy)^n \leq 1-x+e^{-yn}$. To see this, note $x \mapsto (1-xy)^n$ is convex and nonincreasing for fixed y. Also, $x \mapsto 1-x+e^{-yn}$ is linear and nonincreasing for fixed y. Also, $1-y \leq e^{-y}$. [See Lemma 10.5.3.]

We describe the random code.

1. Generate 2^{nR} sequences $\widehat{X}^n(i)$, $i = 1,\ldots,2^{nR}$ i.i.d. from $P_{\widehat{X}}$.
2. Typical set encoding: for each X^n, select i such that $(x^n,\widehat{X}^n(i)) \in A_{d,\epsilon}^{(n)}$ if possible.
   - If there are multiple such i, break ties arbitrarily.
   - If no such i exists, then send $i = 1$. This happens with small probability P_e.

Then,

$$\begin{aligned}
\mathbb{E}d(X^n,\widehat{X}^n(i)) &\leq (1-P_e)(D+\epsilon) + P_e d_{\max} \qquad\qquad d_{\max} \text{ can be relaxed to } \min_{\widehat{x}} \mathbb{E}d(X,\widehat{x}) < \infty\\
&\leq D + \epsilon + P_e d_{\max}.
\end{aligned}$$

It now suffices to show $P_e \to 0$ provided $R > I(X;\widehat{X})$.

$$P((x^n,\widehat{X}^n) \notin A_{d,\epsilon}^{(n)}) = 1 - \sum_{\widehat{x}^n} p(\widehat{x}^n)\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n,\widehat{x}^n)$$

$$P(\nexists i : (x^n,\widehat{X}^n(i)) \in A_{d,\epsilon}^{(n)}) = \left[1 - \sum_{\widehat{x}^n} p(\widehat{x}^n)\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n,\widehat{x}^n)\right]^{2^{nR}} \qquad \text{independence}$$

$$
\begin{aligned}
P_e &= \sum_{x^n} p(x^n)\left[1 - \sum_{\widehat{x}^n} p(\widehat{x}^n)\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n, \widehat{x}^n)\right]^{2^{nR}} \\
&\le \sum_{x^n} p(x^n)\left[1 - \sum_{\widehat{x}^n} p(\widehat{x}^n \mid x^n)2^{-n(I(X;\widehat{X})+3\epsilon)}\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n, \widehat{x}^n)\right]^{2^{nR}} \\
&= \sum_{x^n} p(x^n)\left[1 - \underbrace{2^{-n(I(X;\widehat{X})+3\epsilon)}}_{y}\underbrace{\sum_{\widehat{x}^n} p(\widehat{x}^n \mid x^n)\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n, \widehat{x}^n)}_{x}\right]^{2^{nR}} \\
&\le \sum_{x^n} p(x^n)\left[1 - \sum_{\widehat{x}^n} p(\widehat{x}^n \mid x^n)\mathbf{1}_{A_{d,\epsilon}^{(n)}}(x^n, \widehat{x}^n) + e^{-2^{-n(I(X;\widehat{X})+\epsilon)}2^{nR}}\right] && \text{inequality in "3" above} \\
&= P((X^n, \widehat{X}^n) \notin A_{d,\epsilon}^{(n)}) + e^{-2^{n(R-(I(X;\widehat{X})+3\epsilon))}} \\
&\to 0 + 0 && R > I(X;\widehat{X}) + 3\epsilon
\end{aligned}
$$

□

Recall the rate distortion setup. We give X^n (drawn i.i.d. from some P_X) to an encoder, who then sends nR bits to a decoder, who then outputs $\widehat{X}^n$ which hopefully has distortion $\mathbb{E}d(X^n, \widehat{X}^n) \le D$.

We want to characterize $R(D) = \inf\{R : (R, D) \text{ achievable}\}$, the lowest possible rate at which it is possible to obtain [asymptotically] expected distortion $\le D$.

It turns out that

$$R(D) = \min_{P_{\widehat{X}|X}:\mathbb{E}d(X,\widehat{X})\le D} I(X;\widehat{X}).$$

Last time, we proved achievability: if $R > R(D)$, then there exists a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} \mathbb{E}d(X^n, \widehat{X}^n) \le D$.

Sketch:

1. Fix $P_{\widehat{X}|X}$ such that $\mathbb{E}_{P_X P_{\widehat{X}|X}} d(X, \widehat{X}) \le D$.
2. We picked 2^{nR} sequences $\widehat{X}^n(1), \ldots, \widehat{X}^n(2^{nR})$. For each, there is a distortion ball $\{X^n : d(X^n, \widehat{X}(i)) \le D\}$.
3. If $R > R(D)$, then we have chosen enough $\widehat{X}^n(i)$ so that the set of all corresponding distortion balls is so large that the probability of error is small...

We now turn to proving the converse. First, we need the following lemma.

Lemma 9.4. $D \mapsto \min_{P_{\widehat{X}|X}:\mathbb{E}d(X,\widehat{X})\le D} I(X;\widehat{X})$ is convex.

We proved convexity of the operational definition of $R(D)$ last time by showing the set of achievable pairs (R, D) is convex. However, this lemma asserts convexity of the thing that we have yet to prove is equal to $R(D)$.

Proof. Let $P_{\widehat{X}|X}^{(0)}$ achieve $R(D_0)$, and let $P_{\widehat{X}|X}^{(1)}$ achieve $R(D_1)$.

Define $P_{\widehat{X}|X}^{(\lambda)} = \lambda P_{\widehat{X}|X}^{(0)} + \overline{\lambda} P_{\widehat{X}|X}^{(1)}$.

Since $P_{\widehat{X}|X} \mapsto I(\widehat{X}; X)$ is convex for fixed P_X, we have

$$I_{P_{\widehat{X}|X}^{(\lambda)}}(X;\widehat{X}) \le \lambda I_{P_{\widehat{X}|X}^{(0)}}(X;\widehat{X}) + \overline{\lambda} I_{P_{\widehat{X}|X}^{(1)}}(X;\widehat{X}) = \lambda R(D_0) + \overline{\lambda} R(D_1).$$

$$\mathbb{E}_{P_{\widehat{X}|X}^{(\lambda)}} d(X, \widehat{X}) = \lambda \mathbb{E}_{P_{\widehat{X}|X}^{(0)}} d(X, \widehat{X}) + \overline{\lambda}\mathbb{E}_{P_{\widehat{X}|X}^{(1)}} d(X, \widehat{X}) \le \lambda D_0 + \overline{\lambda} D_1.$$

Thus

$$I_{P^{(\lambda)}_{\widehat{X}|X}}(X;\widehat{X}) \geq \min_{P_{\widehat{X}|X}:\mathbb{E}d(X,\widehat{X})\leq\lambda D_0+\overline{\lambda}D_1} I(X;\widehat{X}) = R(\lambda D_0 + \overline{\lambda}D_1).$$

□

We now prove the converse. We want to prove that if $R < R(D)$, then (R, D) is not achievable by any scheme.

Proof of the converse. Consider a $(2^{nR}, n)$ code with encoder f_n and decoder g_n that achieves $\mathbb{E}d(X^n, \widehat{X}^n) = D$ where $\widehat{X}^n = g_n(f_n(X^n))$. We want to show $R \geq R(D)$.

$$\begin{aligned}
nR &\geq H(\widehat{X}^n) \\
&\geq I(\widehat{X}^n; X^n) \\
&= H(X^n) - H(X^n \mid \widehat{X}^n) \\
&= \sum_{i=1}^n (H(X_i) - H(X_i \mid \widehat{X}^n, X^{i-1}) \\
&\geq \sum_{i=1}^n H(X_i) - H(X_i \mid \widehat{X}_i) && \text{conditioning reduces entropy} \\
&= \sum_{i=1}^n I(X_i; \widehat{X}_i) \\
&\geq \sum_{i=1}^n R(\mathbb{E}d(X_i, \widehat{X}_i)) && \text{def. of } R(D) \\
&\geq nR\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}d(X_i;\widehat{X}_i)\right) && \text{convexity, Jensen} \\
&= nR(D). && \mathbb{E}\frac{1}{n}\sum_{i=1}^n d(X_i;\widehat{X}_i) = \mathbb{E}d(X^n;\widehat{X}^n) = D
\end{aligned}$$

Remark: $nR \geq I(\widehat{X}^n; X^n)$ could be deduced directly by data-processing, since there is a "bottleneck" of nR bits in the model. (?) □

Let us see what happens when the inequalities become tight, in order to characterize a good scheme.

- $nR = H(\widehat{X}_n)$: all 2^{nR} reproductions $\widehat{X}(i)$ are equally likely.
- $H(\widehat{X}^n \mid X^n) = 0$, i.e. $\widehat{X}^n$ is a deterministic function of X^n. [This shows that randomized decoding doesn't help.]
- $H(X_i \mid \widehat{X}^n, X_1, \ldots, X_{i-1}) = H(X_i \mid \widehat{X}_i)$, i.e. $\widehat{X}_i$ is a sufficient statistic for X_i.
- $P_{\widehat{X}_i|X_i} = \operatorname{argmin}_{P_{\widehat{X}|X}:\mathbb{E}d(X,\widehat{X}\leq D} I(X;\widehat{X})$.
- In the application of Jensen, either $R(D)$ is linear (usually isn't), or, in the strictly convex case, $\mathbb{E}d(X_i, \widehat{X}_i) = D$ (exactly the same) for all i.

We now consider **joint source channel coding**. Let V^m be the observation (i.i.d. from P_V). We encode V^m and encodes it into X^n, which gets sent through a discrete memoryless channel (DMC) $P_{Y|X}$. The channel outputs Y^n is then decoded into a reproduction $\widehat{V}^m$ of the original observation. We would like $\mathbb{E}d(V^m, \widehat{V}^m) \leq D$.

Theorem 9.5. Distortion D is achievable if and only if $R(D) \leq BC$ where $B = \frac{n}{m}$ is the bandwith mismatch.

This has a separation result, in that the following scheme is optimal. Do the rate-distortion optimal encoding to $nR(D)$ bits, and these are uniformly distributed. This what the channel likes. Use a channel code at rate C and send it through the channel. Finally, do the corresponding channel decoding and the rate-distortion decoding.

The "if" part is simple: the rate R is lower than the capacity, so everything works.

For the reverse,

$$\begin{aligned} nC &\geq I(X^n; Y^n) && \text{channel coding converse} \\ &\geq I(V^m; \widehat{V}^m) && \text{data processing} \\ &\geq mR(D). && \text{rate distortion converse} \end{aligned}$$

Thus $R(D) \leq BC$.

10 Approximating distributions and entropy

"Approximating Probability Distributions with Dependence Trees" (Chow-Liu 1968)

P is a joint didstribution on n variables $x = (x_1, \ldots, x_n)$. Estimating this is hard (curse of dimensionality). Note

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

where $m_1, \ldots, m_n$ is any permutation of $[n]$.

We want to approximate P by a "second order" distribution This is also known as "tree dependence."

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

where $j(i)$ is the parent of i. Note that these approximations use the same P that we are estimating. So for each tree we have an explicit approximation. We are not approximating P with *any* distribution with a tree structure.

We want to optimize

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

Note $|T_n| = n^{n-2}$.

Note $D(P\|P_t) \geq \frac{\ln 2}{2} \|P - P_t\|_{TV}^2$.

A **maximum weight dependence tree** is a tree t satisfying

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

In other words, if we consider the complete graph on n vertices with edge weights $I(X_i, X_j)$, then the maximum weight spanning tree is this maximum weight dependence tree.

Theorem 10.1. $t^* \in \operatorname{argmin}_{t \in T_n} D(P\|P_t)$ if and only if it is a maximum-weight dependence tree.

Proof.

$$\begin{aligned} D(P\|P_t) &= \sum_x P(x) \log \frac{P(x)}{P_t(x)} \\ &= \sum_x P(x) \log P(x) - \sum_x P(x) \sum_{i=1}^{n} \log P(x_i \mid x_{j(i)}) \\ &= -H(X) - \sum_x P(x) \sum_{i=1}^{n} \log \frac{P(x_i, x_{j(i)})}{P(x_{j(i)})P(x_i)} - \sum_x P(x) \sum_{i=1}^{n} \log P(x_i) \\ &= \underbrace{-H(X) + \sum_{i=1}^{n} H(X_i)}_{\text{no dependence on } t} - \sum_{i=1}^{n} I(X_i; X_{j(i)}). \end{aligned}$$

So $\operatorname{argmin}_t D(P\|P_t) = \operatorname{argmax}_t \sum_{i=1}^n I(X_i; X_{j(i)})$. □

If we want to do this approximation, we just need to estimate the $O(n^2)$ mutual informations, rather than the $O(2^n)$ probabilities (if x_i are binary).

Maximum likelihood estimator of the true tree is the plug-in estimator. (Estimate mutual informations using data and take the empirical maximum spanning tree.)

Why relative entropy ends up being nice? Chain rule / factorization is one possibility.

Our problem is as follows. We have $X^{(1)}, \ldots, X^{(m)}$ and we want to get an estimate of the dependence tree. We estimate the mutual informations $I(X_i, X_j)$ and find the empirical max weight tree.

Recall $I(X;Y) = H(X) + H(Y) - H(X,Y)$. So it suffices to estimate entropies.

How do we estimate $H(P)$ given n i.i.d. samples drawn from P?

Classical statistics. Suppose $|\mathcal{X}| = S$ is fixed. Find the optimal estimator of $H(P)$ as $n \to \infty$. Let P_n denote the empirical distribution, then

$$H(P_n) = -\sum_i \widehat{p}_i \log \widehat{p}_i$$

where $\widehat{p}_i$ is the relative frequency of symbol i in the data.

$H(P_n)$ is the MLE for $H(P)$, and it is asymptotically efficient (asymptotically, attains equality in the Cramer-Rao bound) by Hajek-Le Cam theory.

What about non-asymptotics? What if n is not "huge" relative to the alphabet size S?

Decision theoretic framework. For an estimator $\widehat{H}_n$, the worst-case risk is

$$R_n^{\max}(\widehat{H}_n) = \sup_{P \in \mathcal{M}_S} \mathbb{E}_P (H(P) - \widehat{H}_n)^2,$$

and the minimax risk is

$$\inf_{\widehat{H}_n} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P (H(P) - \widehat{H}_n)^2.$$

Classical asymptotics: for the plug-in estimator,

$$\mathbb{E}_P (H(P) - H(P_n))^2 \sim \frac{\operatorname{Var}(-\log P(X))}{n}$$

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

Does $n \gg (\log S)^2$ imply consistency?

No. Bias-variance decomposition:

$$\mathbb{E}_P (H(P) - \widehat{H})^2 = (\mathbb{E}[\widehat{H}] - H(P))^2 + \operatorname{Var}_P(\widehat{H})$$

Jiao Venkat Han Weissman (2014): for the plug-in estimator,

$$R_n^{\max}(H(P_n)) \asymp \underbrace{\frac{S^2}{n^2}}_{\text{bias squared}} + \underbrace{\frac{(\log S)^2}{n}}_{\text{variance}}$$

If $n \gg S$ (e.g. $n \to \infty$ while S fixed) then the bias term is small and we get the classical asymptotics. Otherwise, bias term becomes important. So, consistency of MLE is equivalent to $n = \Theta(S)$.

Next time, we show we can do better.

Recall we are trying to estimate $H(P)$ for some distribution P, using i.i.d. samples X^n. The MLE is $H(P_n)$ where P_n is the empirical distribution.

From classical statistics,

$$\mathbb{E}_P (H(P) - H(P_n))^2 \sim \frac{\operatorname{Var}(-\log P(X))}{n}$$

as $n \to \infty$. If S is the support size, this suggests the sample complexity is $\Theta((\ln S)^2)$. However, this is only valid in the asymptotic regime $n \to \infty$ while support size is fixed.

If the support as size S,

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P(H(P)-H(P_n))^2 \asymp \frac{S^2}{n^2}+\frac{(\ln S)^2}{n}.$$

So the sample complexity of MLE is actually $\Theta(S)$. If n is not large enough, the bias term (first term) is too large.

There is a phase transition: if sample complexity is $\gtrsim S$ then the risk is nearly zero, if it is less, than the risk is high.

Can we do better than the MLE? Valiant and Valiant showed that the [minimax?] phase transition for entropy estimation occurs instead at $\Theta(S/\ln S)$.

We get an "effective sample size enlargement phenomena." This result implies that the risk with n samples has the same error as the MLE estimator with $n\log n$ samples.

Note that entropy is separable: $H(P)=-\sum_i p_i\log p_i=\sum_i f(p_i)$ where $f(x)=-x\log x$.

Recall the issue with the MLE is large bias when n is not large enough.

The plug-in estimator is $H(P_n)=-\sum\widehat{p}_i\log\widehat{p}_i$. Consider the function $f(x)=-x\log x$. If $\widehat{p}_i\approx p_i$ and p_i is near $1/2$, then $f(\widehat{p}_i)\approx f(p_i)$ because the slope of f is low. However, near zero, f has infinite slope, so $f(\widehat{p}_i)$ and $f(p_i)$ differ greatly.

To fix this, we divide $[0,1]$ into a smooth regime $(\log n/n,1]$ and a non-smooth regime $[0,\log n/n)$.

For the smooth regime we have a bias-corrected estimate $f(\widehat{p}_i)-\frac{1}{2n}f''(\widehat{p}_i)\widehat{p}_i(1-\widehat{p}_i)$.

For the non-smooth regime $[0,\log n/n)$, use the best (sup norm) polynomial approximation of f order $\log n$.

Note that this estimation procedure does not depend on S.

We compare the L^2 rates:

$$\begin{aligned}\text{minimax}:&\frac{S^2}{(n\log n)^2}+\frac{(\ln S)^2}{n}\\ \text{MLE}:&\frac{S^2}{n^2}+\frac{(\ln S)^2}{n}\end{aligned}$$

11 Computing rate distortion and channel capacity

Alternating minimization algorithm.

Example: Suppose we have disjoint sets A,B and we want to find $\min_{a\in A,b\in B}\|a-b\|^2$. We maintain a current a_t and b_t, and then update $a_{t+1}=\operatorname{argmin}_{a\in A}\|a-b_t\|^2$ and $b_{t+1}=\operatorname{argmin}_{b\in B}\|a_{t+1}-b\|^2$. This is guaranteed to converge to the minimum for well-behaved distance functions and convex A and B.

For us, relative entropy is a good distance function.

Let A be the joint distributions with marginal P_X and expected distortion $\le D$.

$$A:=\left\{Q_{X,\widehat{X}}:\mathbb{E}d(X,\widehat{X})\le D,\sum_{\widehat{x}}Q(x,\widehat{x})=P_X(x)\right\}$$

$$\begin{aligned}R(D)&=\min_{Q\in A}I(X;\widehat{X})\\&=\min_{Q\in A}D(Q_{X,\widehat{X}}\|Q_XQ_{\widehat{X}})\\&=\min_{Q\in A}D(Q_{X,\widehat{X}}\|P_XQ_{\widehat{X}}).\end{aligned}$$

Lemma 11.1.

$$R(D)=\min_{Q\in A}\min_{R_{\widehat{X}}}D(Q_{X,\widehat{X}}\|P_XR_{\widehat{X}}).$$

Proof.

$$D(Q_{X,\widehat{X}}\|P_XR_{\widehat{X}})-D(Q_{X,\widehat{X}}\|P_XQ_{\widehat{X}})=D(Q_{\widehat{X}}\|R_{\widehat{X}})\ge 0.$$

□

To implement the alternating minimization algorithm, we need to solve two problems.

- Given $Q_{X,\widehat{X}}$, find $R_{\widehat{X}}$ that minimizes $D(Q_{X,\widehat{X}} \| P_X R_{\widehat{X}})$. The above lemma implies $R_{\widehat{X}} = Q_{\widehat{X}}$.
- Given $R_{\widehat{X}}$, find $Q_{X,\widehat{X}} \in A$ that minimizes $D(Q_{X,\widehat{X}} \| P_X R_{\widehat{X}})$. Since the marginal $Q_{\widehat{X}}$ is fixed, we just need to find $Q_{\widehat{X}|X}$. The following lemma shows there is a closed form expression.

Lemma 11.2. The minimizing conditional distribution has the form

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

where λ is such that $\mathbb{E}_Q d(X, \widehat{X}) = D$.

Proof. Lagrange multipliers.

$$J(Q_{\widehat{X}|X}) = D(Q_{\widehat{X}|X} P_X \| R_{\widehat{X}} P_X) + \lambda_1 \mathbb{E}_Q d(X, \widehat{X}) + \lambda_2 \sum_{x,\widehat{x}} P_X(x) Q_{\widehat{X}|X}(\widehat{x} \mid x)$$

$$\frac{\partial}{\partial Q_{\widehat{X}|X}(\widehat{x} \mid x)} J(Q_{\widehat{X}|X}) = P_X(x) \log \frac{Q_{\widehat{X}|X}(\widehat{x} \mid x)}{R_{\widehat{X}}(\widehat{x})} + P_X(x) + \lambda_1 P_X(x) d(x, \widehat{x}) + \lambda_2 P_X(x)$$

□

Proof of second part of Lemma 10.8.1.

$$\begin{aligned}
\sum_{x,y} p(x) p(y \mid x) \left(\log \frac{r^*(x \mid y)}{p(x)} - \log \frac{r(x \mid y)}{p(x)} \right) &= \sum_{x,y} p(x) p(y \mid x) \log \frac{r^*(x, y)}{r(x \mid y)} \\
&= p(y) \sum_{x,y} r^*(x \mid y) \log \frac{r^*(x \mid y)}{r(x \mid y)} \\
&= p(y) D(r^*(x \mid y) \| r(x \mid y)) \\
&\geq 0.
\end{aligned}$$

Similarly, there is an alternating maximization procedure for channel capacity.

$$\begin{aligned}
C &= \max_{P_X} I(X; Y) \\
&= \max_{P_X} D(P_{XY} \| P_X P_Y) \\
&= \max_{Q_{X|Y}} \max_{R_X} \sum_{x,y} R_X(x) P_{Y|X}(y \mid x) \log \frac{Q_{X|Y}(x \mid y)}{R_X(x)}
\end{aligned}$$

For any R_X,

$$Q^*_{X|Y}(x \mid y) := \frac{R_X(x) P_{Y|X}(y \mid x)}{\sum_{x'} R_X(x') P_{Y|X}(y \mid x')}.$$

For any $Q_{X|Y}$,

$$R^*_X(x) = \frac{\prod_y Q_{X|Y}(x \mid y)^{P_{Y|X}(y|x)}}{\sum_{x'} \prod_y Q_{X|Y}(x' \mid y)^{P_{Y|X}(y|x')}}$$

12 Information theory and statistics

12.1 Theory of types

Let $X_1, \dots, X_n$ be a sequence of symbols from $\mathcal{X} = \{a_i\}$.
The **type** of X^n, denoted P_{X^n}, is the empirical distribution associated to $X^n = (X_1, \dots, X_n)$.
$\mathcal{P}_n$ denotes the set of types with denominator n, i.e., the possible types associated with a sample of size n.

Example 12.1. If $\mathcal{X} = \{0, 1\}$,

$$\mathcal{P}_n := \left\{ \left(\frac{k}{n}, \frac{n-k}{n} \right) : 0 \le k \le n \right\}$$

■

The **type class** of $P \in \mathcal{P}_n$ is

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

Example 12.2. If $P = (3/8, 5/8)$, then $T(P)$ are all $\binom{8}{3}$ binary vectors of length 8 with exactly three zeros. ■

Theorem 12.3.

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

Proof. There are $n+1$ choices $\{0, 1, \dots, n\}$ for each numerator. □

[The above bound is very crude, but it is good enough because we will be comparing this to things that grow exponentially in n.]

Consequently, the number of type classes is polynomial in n.

Theorem 12.4. Let $X_1, \dots, X_n \sim Q$ be i.i.d. Then the probability of X^n is

$$Q^n(X^n) = 2^{-n(H(P_{X^n}) + D(P_{X^n} \| Q))}.$$

It makes sense that the probability depends only on the type. (Permuting does not affect the empirical distribution.)
Recall $Q^n(X^n) \approx 2^{-nH(Q)}$ for typical X^n and contrast this with the statement of the theorem.

Proof.

$$\begin{aligned} 2^{-n(H(P_{X^n}) + D(P_{X^n} \| Q))} &= 2^{-n\left(\sum_{a \in \mathcal{X}} P_{X^n}(a) \log \frac{1}{P_{X^n}(a)} + \sum_{a \in \mathcal{X}} P_{X^n}(a) \log \frac{P_{X^n}(a)}{Q(a)}\right)} \\ &= 2^{n \sum_{a \in \mathcal{X}} P_{X^n}(a) \log Q(a)} \\ &= \prod_{a \in \mathcal{X}} Q(a)^{n P_{X^n}(a)} \\ &= Q^n(X^n), \end{aligned}$$

where we note $nP_{X^n}(a)$ is the number of times a appears in the sample. □

Theorem 12.5.

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

That is, the type class of P has about $2^{nH(P)}$ sequences.

This is a more precise notion than typical sets. Both shows how entropy is some notion of volume.

Proof.

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

The last equality is due to

$$2^{-nH(P)} = 2^{n \sum_a P(a) \log P(a)} = \prod_a P(a)^{nP(a)}.$$

[Alternatively, apply the previous theorem with $Q = P_{X^n}$.]

We now prove the lower bound. We will assume

$$P^n(T(P)) \geq P^n(T(\widehat{P})), \forall \widehat{P} \in \mathcal{P}_n$$

and prove it later. Intuitively, it states that under a particular probability distribution, the type with maximum probability is the original distribution.

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

It remains to prove the "maximum likelihood" result.

$$\begin{aligned}\frac{P^n(T(P))}{P^n(T(\widehat{P}))} &= \frac{|T(P)|}{|T(\widehat{P})|} \cdot \frac{\prod_a P(a)^{nP(n)}}{\prod_a P(a)^{n\widehat{P}(a)}} \\ &= \frac{\binom{n}{nP(a_1),\ldots,nP(a_{|\mathcal{X}|})}}{\binom{n}{n\widehat{P}(a_1),\ldots,n\widehat{P}(a_{|\mathcal{X}|})}} \cdot \prod_a P(a)^{n(P(a)-\widehat{P}(a))} \\ &= \frac{(n\widehat{P}(a_1))! \cdots (n\widehat{P}(a_{|\mathcal{X}|}))!}{(nP(a_1))! \cdots (nP(a_{|\mathcal{X}|}))!} \cdot \prod_a P(a)^{n(P(a)-\widehat{P}(a))} \\ &\geq \prod_a (nP(a))^{n(\widehat{P}(a)-P(a))} P(a)^{n(P(a)-\widehat{P}(a))} && \frac{m!}{n!} \geq n^{m-n} \\ &= \prod_a n^{n(\widehat{P}(n)-P(a))} \\ &= n^{n \sum_a (\widehat{P}(a)-P(a))} \\ &= 1.\end{aligned}$$

□

Theorem 12.6. For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of type class $T(P)$ under Q is

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

The probability of observing some empirical distribution under Q is exponentially small in the relative entropy.

Proof.

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

Applying the previous theorem finishes the proof. □

In summary,

- $|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}$. (Bound on number of types.)
- $Q^n(x^n) = 2^{-n(H(P)+D(P\|Q))}$ for $x^n \in T(P)$.
- $|T(P)| \doteq 2^{nH(P)}$
- $Q^n(T(P)) \doteq 2^{-nD(P\|Q)}$

Theorem 12.7. Let $X_1, X_2, \ldots$ be i.i.d. from P. Then

$$\mathbb{P}(D(P_{X^n}\|P) > \epsilon) \le 2^{-n\left(\epsilon - \frac{1}{n}|\mathcal{X}|\log(n+1)\right)}$$

Note that the right-hand side does not depend on P.
The Borel-Cantelli theorem implies that if $\sum_n \mathbb{P}(D(P_{X^n}\|P) > \epsilon) < \infty$, then $D(P_{X^n}\|P) \to 0$ almost surely.

Given $\epsilon > 0$, let

$$T_Q^\epsilon := \{x^n : D(P_{X^n}\|Q) \le \epsilon\}.$$

$$\begin{aligned} \mathbb{P}_{Q^n}\{X^n : D(P_{X^n}\|Q) > \epsilon\} &= 1 - Q^n(T_Q^\epsilon) \\ &= \sum_{P \in \mathcal{P}_n : D(P\|Q) > \epsilon} Q^n(T(P)) \\ &\le \sum_{P \in \mathcal{P}_n : D(P\|Q) > \epsilon} 2^{-nD(P\|Q)} \\ &\le (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \\ &= 2^{-n\left(\epsilon - \frac{|\mathcal{X}|\log(n+1)}{n}\right)} \end{aligned}$$

This is a law of large numbers: the probability of getting a sample whose empirical distribution is far from Q in relative entropy is exponentially small. Applying Borel-Cantelli implies

$$D(P_{X^n}\|Q) \to 0, \quad \text{almost surely.}$$

This is a strengthening of the law of large numbers.

Note that relative entropy controls L^1 distance between measures. On finite-dimensional spaces, all norms are equivalent, so up to constants relative entropy controls all norms (in finite dimensions). More generally, relative entropy controls many transportation distances.

12.2 Large deviations

Let $X_1, X_2, \ldots \sim Q$ be i.i.d. on a finite alphabet $\mathcal{X}$.
The weak law of large numbers states

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > \mathbb{E}X_1 + \epsilon\right) \to 0.$$

The proof using Chebychev's inequality actually gives us a rate

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > \mathbb{E}X_1 + \epsilon\right) \leq \frac{\mathrm{Var}(X)}{n^2\epsilon^2}.$$

We usually rewrite the left-hand side as

$$P\left(\sum_{i=1}^{n} X_i > n\mathbb{E}X_1 + n\epsilon\right) \doteq 2^{-nE},$$

where $n\epsilon$ is called the **large deviation**. What is interesting is that E is an exponent that we can compute explicitly, and the upper bound is tight up to a constant.

Example 12.8. Let $X_i \sim \mathrm{Ber}(p)$ and $Q = \mathrm{Ber}(p)$. Note that $\frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}_{X\sim P_{X^n}} X = P_{X^n}(1)$ (the proportion of 1s under the empirical distribution).

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \epsilon\right) &= \sum_{P\in\mathcal{P}_n : P(1)\geq p+\epsilon} Q^n(T(P)) \\ &\in \left[\frac{|\mathcal{P}_n|}{(n+1)^{\mathcal{X}}} 2^{-n\min D(P\|Q)}, |\mathcal{P}_n| 2^{-n\min D(P\|Q)}\right]\end{aligned}$$

where the minimum is over the same types in the sum.

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \epsilon\right) \doteq 2^{-n\min D(P\|Q)}$$

■

This example is a special case of the following theorem, with E being the collection of distributions on $\{0,1\}$ with expectation $\geq p + \epsilon$.

Theorem 12.9 (Sanov's theorem)**.** Let $X_1, X_2, \ldots \sim Q$ be i.i.d. and let E be a collection of probability distributions on $\mathcal{X}$. Then the probability that the empirical distribution P_{X^n} lies in E is

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

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

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

that is, the lower bound matches the upper bound.

Common example of the collection E is $E := \left[P : \sum_{x\in\mathcal{X}} g(x)P(x) \geq \alpha\right]$, that is, the set of distributions whose such that $\mathbb{E}_{X\sim P} g(X) \geq \alpha$. If $g(x) = x^k$, this is a moment constraint. We could also have many constraints (g_j and α_j for $j = 1, \ldots, J$).

Proof.

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

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

We need to find a sequence $\{P_n : P_n \in E \cap \mathcal{P}_n\}_{n \ge 1}$ such that $D(P_n\|Q) \to D(P^*\|Q)$. If E has nonempty interior and E is the closure of its interior, then we can approximate any interior point by a sequence of types, and this is possible. □

We review Sanov's theorem. We consider the collection of probability distributions on $\mathcal{X}$ and observe $X_1, \dots, X_n \sim Q$ for one particular distribution. Let E be a collection of other distributions, e.g., set of distributions with expected value ≥ 0.8. We want to understand the probability that the empirical distribution P_{X^n} is in E. Sanov's theorem implies that this probability exponentially small with exponent $\min_{P \in E} D(P\|Q)$.

If $P^* := \operatorname{argmin}_{P \in E} D(P\|Q)$, then we get the lower bound $Q^n(T(P^*)) \ge \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P^*\|Q)}$ for free. The "closure of the interior" condition allows use to use denseness of types to extend to the case where P^* is not a type.

Example 12.10. Suppose we have a fair coin. What is the probability of ≥ 700 heads in 1000 tosses? Let $E = \{P : \mathbb{E}_{X \sim P} X \ge 0.7\}$. Why does this make sense? If the observations X^n have ≥ 700 heads then $P_{X^n} \in E$ (here $n = 1000$). Note that Q is the fair distribution. Sanov's theorem implies

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

More precisely,

$$2^{-138.9} = 2^{-n(0.119+0.0199)} \le \mathbb{P}(\ge 700 \text{ heads}) \le 2^{-n(0.119-0.0199)} = 2^{-99.1}.$$

Again, $n = 1000$ (but the same argument works for any n). The upper and lower bounds differ only in an exponential factor of $\log n/n$; as $n \to \infty$ these are very close. ■

To compute $P^* = \operatorname{argmin}_{P \in E} D(P\|Q)$ where E is convex, this becomes a convex optimization problem: use Lagrange multipliers.

The more general version of Sanov's theorem (continuous distributions, etc.) is as follows.

Theorem 12.11 (Sanov's theorem)**.**

$$\begin{aligned} -\inf_{P \in \operatorname{int}(E)} D(P\|Q) &\le \liminf_{n \to \infty} \frac{1}{n} \log Q^n(E) \\ &\le \limsup_{n \to \infty} \frac{1}{n} \log Q^n(E) \\ &\le -\inf_{P \in \operatorname{cl}(E)} D(P\|Q). \end{aligned}$$

12.3 Conditional limit theorem

Suppose I am manufacturing bolts, each of which is supposed to nominally weight 10 grams. I find a batch of 1000 bolts that weighs ≥ 10.5 kilograms. What is the probability that any given bolt weights 11 grams?

[What does the bulk measurement tell us about the marginal distributions of the individual measurements?]

Theorem 12.12 (Conditional limit theorem)**.** Suppose $X_1, X_2, \ldots \sim Q$ are i.i.d. and we observe $P_{X^n} \in E$ with $Q \notin E$ and E is closed and convex.

$$\mathbb{P}(X_1 = a \mid P_{X^n} \in E) \xrightarrow{p} P^*(a),$$

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

We need two intermediate results along the way.

Theorem 12.13 (Pythagorean theorem)**.** For $E \subset \mathcal{P}(\mathcal{X})$ closed and convex, and $Q \notin E$, let $P^* := \operatorname{argmin}_{P \in E} D(P\|Q)$. Then,

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

for all $P \in E$.

Proof. Let $P \in E$ and define $P_\lambda = \lambda P + \overline{\lambda} P^*$. By definition of P^*, we have $\frac{d}{d\lambda} D(P_\lambda\|Q) \geq 0$ at $\lambda = 0$.

$$\begin{aligned}
D(P_\lambda\|Q) &= \sum_x P_\lambda(x) \log \frac{P_\lambda(x)}{Q(x)} \\
\frac{d}{d\lambda} D(P_\lambda\|Q) &= \sum_x \left[(P(x) - P^*(x)) \log \frac{P_\lambda(x)}{Q(x)} + P(x) - P^*(x) \right] \\
&= \sum_x (P(x) - P^*(x)) \log \frac{P_\lambda(x)}{Q(x)} \\
0 \leq \frac{d}{d\lambda} D(P_\lambda\|Q)\Big|_{\lambda=0} &= \sum_x (P(x) - P^*(x)) \log \frac{P^*(x)}{Q(x)} \\
&= \sum_x P(x) \log \frac{P^*(x)}{Q(x)} - D(P^*\|Q) \\
&= D(P\|Q) - D(P\|P^*) - D(P^*\|Q).
\end{aligned}$$

□

Theorem 12.14 (Pinsker's inequality)**.**

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

Note that for $A = \{x : P(x) \geq Q(x)\}$,

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

Proof. For binary distributions, one can prove the following (exercise):

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

The data processing inequality for relative entropy: if $P' = P_{Y|X}P$ and $Q' = P_{Y|X}Q$, then

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

Define a channel $Y = \mathbf{1}_{\{X \in A\}}$. Then

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

□

Intuition for the conditional limit theorem: P^* completely dominates the behavior of the marginal.

Proof of conditional limit theorem. Let $S_t := \{P \in \mathcal{P}(\mathcal{X}) : D(P\|Q) \leq t\}$. This is a convex set.

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

Let $A := S_{D^*+2\delta} \cap E$ and $B := E \setminus A = E \setminus S_{D^*+2\delta}$.

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

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

$$\mathbb{P}(P_{X^n} \in B \mid P_{X^n \in E}) = \frac{Q^n(B \cap E)}{Q^n(E)} \leq \frac{Q^n(B)}{Q^n(A)} \leq \frac{(n+1)^{|\mathcal{X}|}2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}}2^{-n(D^*+\delta)}} = (n+1)^{2|\mathcal{X}|}2^{-n\delta} \to 0.$$

[So, the probability that our empirical distribution is outside a KL-ball around P^* (given it lies in E) vanishes.] Thus,

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

By the Pythagorean inequality, for $P \in A$ we have

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

so $D(P\|P^*) \leq 2\delta$. This combined with Pinsker's inequality implies

$$\mathbb{P}(P_{X^n} \in A \mid P_{X^n} \in E) \leq \mathbb{P}(D(P_{X^n}\|P^*) \leq 2\delta \mid P_{X^n} \in E) \leq \mathbb{P}(\|P_{X^n} - P^*\|_1 \leq \delta' \mid P_{X^n} \in E),$$

and by our earlier work, these three quantities tend to 1 as $n \to \infty$. Consequently

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

□

Aside: after proving Sanov's theorem, we proved $D(P_{X^n}\|P) \to 0$ almost surely which implies $\|P_{X^n} - P\|_1 \to 0$ almost surely, by Pinsker's inequality.

12.4 Fisher information and Cramer-Rao lower bound

Let $f(x;\theta)$ be a family of densities indexed by θ. For example, the location family is $f(x;\theta) = f(x-\theta)$ for some f.

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

For example, $X_i \sim \mathcal{N}(\theta, 1)$ i.i.d. and $T(X^n) = \frac{1}{n}\sum_{i=1}^n X_i$.

An estimator is unbiased if $\mathbb{E}_\theta T(X^n) = \theta$.

The **Cramer-Rao bound** states that the variance of an unbiased estimator is lower bounded by $1/J(\theta)$.

Example 12.15. Let $f(x;\theta) := f(x-\theta)$. Then $J(\theta) = \int \frac{f'(x)^2}{f(x)}\,dx$. This is a measure of curvature/smoothness. If f is very spread out, it is hard to estimate θ; indeed then $1/J(\theta)$ will be large. ■

We will see later that there is a finer inequality.

$$\frac{1}{J(X)} \le \frac{1}{2\pi e} 2^{2h(X)} \le \mathrm{Var}(X).$$

13 Entropy methods in mathematics

13.1 Fisher information and entropy

Let $u(t,x)$ denote the temperature at time x at time t, where $x \in \mathbb{R}^n$. The heat equation is

$$\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}u(t,x).$$

Consider the initial condition $u(0,x) = \delta(x)$ (all heat starts at 0). Then a solution is

$$u_0(t,x) = (2\pi t)^{-n/2} e^{-|x|^2/2t}.$$

That is, at time t, the temperature profile is the Gaussian density with variance t.

More generally, if the initial condition is $u(0,x) = f(x)$, then a solution is the convolution $u(t,x) = \int f(s)u_0(t,x-s)\,ds$ where u_0 is the Gaussian kernel above. Note that if we integrate over x, we get the "total energy" which is conserved (constant in t). This is easy to see in the special case where f is a density (in x), in which case the convolution will also be a density, which integrates (over x) to 1, which is constant in t.

Let us focus on this special case. Let $X \sim f$ and $Z \sim N(0,I)$. Then $u(t,x) = f_t(x)$ where f_t is the density of $X + \sqrt{t}Z$. [Convolution of densities is density of sum.]

We claim $h(X+\sqrt{t}Z)$ is nondecreasing in t. [This matches our intuition from the second law of thermodynamics.]

$$h(X+\sqrt{t}Z) \ge h(X+\sqrt{t}Z \mid Z) = h(X).$$

This implies

$$h(X+\sqrt{t+t'}Z) = h(X+\sqrt{t}Z+\sqrt{t'}Z') \ge h(X+\sqrt{t}Z)$$

which implies our claim.

Amazingly, we not only know it is increasing in t, but we have a formula for the rate of increase.

Proposition 13.1 (de Bruijn's identity)**.**

$$\frac{d}{dt}h(X+\sqrt{t}Z) = \frac{1}{2}J(X+\sqrt{t}Z).$$

That is,

$$\frac{d}{dt}h(f_t) = \frac{1}{2}J(f_t).$$

We will define the Fisher information J now. Given a parametric family of densities $\{f(x;\theta)\}$ parameterized by θ, the Fisher information at θ is

$$J(\theta) = \mathop{\mathbb{E}}_{X\sim f(x;\theta)}\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2.$$

We will focus on location families where $f(x;\theta) := f_X(x-\theta)$ for some fixed density f_X. In this case, the Fisher information is

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

which is free of θ. We then use the compact notation $J(\theta) = J(f) = J(X)$ to denote the Fisher information associated with this density. This is the Fisher information that we will consider from now on.

In n dimensions, this generalizes to

$$J(f) = J(X) = \int \frac{|\nabla f(x)|^2}{f(x)}\,dx = 4\int\left|\nabla\sqrt{f(x)}\right|^2 dx\,.$$

The last expression shows how the Fisher information corresponds to the smoothness of f (actually, of $\sqrt{f}$).

We now prove de Bruijn's identity.

Proof.

$$\begin{aligned}
-\frac{d}{dt}h(f_t) &= \frac{d}{dt}\int f_t(x)\log f_t(x)\,dx \\
&= \int(\log f_t(x))\frac{\partial}{\partial t}f_t(x)\,dx + \int\frac{\partial}{\partial t}f_t(x)\,dx \\
&= \int(\log f_t(x))\frac{1}{2}\sum_{i=1}^n\frac{\partial^2}{\partial x_i^2}f_t(x)\,dx + \underbrace{\frac{d}{dt}\int f_t(x)\,dx}_{=\frac{d}{dt}1=0} \\
&= \frac{1}{2}\sum_{i=1}^n\int(\log f_t(x))\frac{\partial^2}{\partial x_i^2}f_t(x)\,dx\,.
\end{aligned}$$

Assuming $(\log f_t(x))\frac{\partial}{\partial x_i}f_t(x)\to 0$ as $|x|\to\infty$, integration by parts for each summand indexed by i (with respect to dx_i) gives

$$\begin{aligned}
-\frac{d}{dt}h(f_t) &= \frac{1}{2}\sum_{i=1}^n\int(\log f_t(x))\frac{\partial^2}{\partial x_i^2}f_t(x)\,dx \\
&= -\frac{1}{2}\sum_{i=1}^n\int\frac{\left(\frac{\partial}{\partial x_i}f_t(x)\right)^2}{f_t(x)}\,dx \\
&= -\frac{1}{2}\int\frac{|\nabla f_t(x)|^2}{f_t(x)}\,dx \\
&= -\frac{1}{2}J(f_t).
\end{aligned}$$

□

Note that nonnegativity of Fisher information also shows that $h(f_t)$ is nondecreasing in t.

We will use de Bruijn's identity to prove an uncertainty principle for entropy and Fisher information.

The entropy power inequality gives

$$\begin{aligned}
e^{\frac{2}{n}h(X+\sqrt{t}Z)} &\geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(\sqrt{t}Z)} = e^{\frac{2}{n}h(X)} + 2\pi et \\
\frac{e^{\frac{2}{n}h(X+\sqrt{t}Z)} - e^{\frac{2}{n}h(X)}}{t} &\geq 2\pi e.
\end{aligned}$$

Taking $t \to 0$ makes the left-hand side equal to

$$\frac{d}{dt} e^{\frac{2}{n} h(X+\sqrt{t}Z)} \Big|_{t=0} = \frac{1}{n} J(X) e^{\frac{2}{n} h(X)},$$

by de Bruijn's identity. Thus, we arrive at the following.

Proposition 13.2 (Stam's inequality)**.**

$$J(X) e^{\frac{2}{n} h(X)} \geq 2\pi e n.$$

This is an uncertainty principle: product of two uncertanties is greater than some constant.

Note that in dimension $n = 1$, we have $h(X) \leq \frac{1}{2} \log[2\pi e \operatorname{Var}(X)]$, so we recover the Cramer-Rao bound

$$J(X) \operatorname{Var}(X) \geq 1.$$

13.2 The logarithmic Sobolev inequality

Let ϕ be the density of $Z \sim N(0, I)$. Let $\phi(x)\, dx = d\gamma$. We have

$$D(X\|Z) = \int f(x) \log \frac{f(x)}{\phi(x)} = \int \frac{f}{\phi} \log \frac{f}{\phi}\, d\gamma$$

The relative Fisher information is

$$I(X\|Z) = \int f(x) \left| \nabla \log \frac{f(x)}{\phi(x)} \right|^2 dx = \int \frac{f}{\phi} \left| \nabla \log \frac{f}{\phi} \right|^2 d\gamma\,.$$

Recalling $h(X) = -\int f \log f\, dx$ and $J(X) = \int f |\nabla \log f|^2\, dx$, we see that the above two quantities are parallel analogues of entropy and Fisher information.

The above two quantities can be simplified to be

$$\begin{aligned} D(X\|Z) &= \frac{n}{2} \log(2\pi e) + \frac{1}{2} \mathbb{E}|X|^2 - \frac{n}{2} - h(X) \\ I(X\|Z) &= J(X) - 2n + \mathbb{E}|X|^2. \end{aligned}$$

Using the bound $\log x \leq x - 1$, we have

$$\log 2\pi e \leq \log\left(\frac{1}{n} J(X) \right) + \frac{2}{n} h(X) \leq \frac{1}{n} J(X) - 1.$$

Combining this bound with the above implies the following.

Theorem 13.3 (Log Sobolev inequality, information-theoretic form)**.**

$$\frac{1}{2} I(X\|Z) \geq D(X\|Z).$$

$$\text{EPI} \overset{\text{de Bruijn}}{\Longrightarrow} J(X) e^{\frac{2}{n} H(X)} \geq 2\pi e n \iff D(X\|Z) \leq \frac{1}{2} I(X\|Z)$$

We only showed the forward implication of the last "if and only if." The reverse is simple too.

Note that the last result is dimension free.

Let $g^2 := f/\phi$. We can reformulate the last result.

$$2 \int |\nabla g|^2\, d\gamma \geq \int g^2 \log g^2\, d\gamma\,.$$

Noting that $\int g^2\, d\gamma = 1$, we can write

$$2 \int |\nabla g|^2\, d\gamma \geq \int g^2 \log g^2\, d\gamma - \int g^2\, d\gamma \log \int g^2\, d\gamma\,.$$

This inequality still holds if we scale g by a constant. (Log terms will cancel.)

Theorem 13.4 (Log Sobolev inequality for Gaussian measure)**.** For "smooth" g,

$$2\int |\nabla g|^2\, d\gamma \geq \int g^2 \log g^2\, d\gamma - \int g^2\, d\gamma \log \int g^2\, d\gamma =: \mathrm{Ent}_\gamma(g^2).$$

This is equivalent to the previous formulation of the log Sobolev inequality.

EPI $\implies$ Fisher information-entropy uncertainty principle $\iff$ LSI (info. th.) $\iff$ LSI (functional form)

13.3 Concentration of measure

$F : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz (denoted $\|F\|_{Lip} \leq L$) if $|F(x) - F(y)| \leq L|x - y|$, for all x and y.

Theorem 13.5 (Borell's inequality)**.** Let $Z \sim \mathcal{N}(0, I)$. If $\|F\|_{Lip} \leq L$, then

$$\mathbb{P}(F(Z) \geq \mathbb{E}[F(Z)] + r) \leq e^{-\frac{r^2}{2L^2}}.$$

Consider $U \sim N(0, 1)$. We have $\mathbb{P}(U \geq r) \leq e^{-r^2/2}$. So, the theorem states that under a Lipschitz function, the tail behavior is still the same.

Without loss of generality suppose $L = 1$. We consider $g^2(x) := e^{\lambda F(x) - \lambda^2/2}$ (and assume $\int F\, d\gamma = 0$) and plug it into the LSI. We have $\nabla g(x) = \frac{\lambda}{2}(\nabla F(x))e^{\lambda/2F(x) - \lambda^2/4}$, so

$$\begin{aligned} 2\int |\nabla g|^2\, d\gamma &= \frac{\lambda^2}{2}\int |\nabla F|^2 e^{\lambda F - \lambda^2/2}\, d\gamma \\ &\leq \frac{\lambda^2}{2}\int e^{\lambda F - \lambda^2/2}\, d\gamma\,. \end{aligned}$$

Let $\Lambda(\lambda) := \int e^{\lambda F - \lambda^2/2}\, d\gamma$. The LSI implies

$$\begin{aligned} \frac{\lambda^2}{2}\int e^{\lambda F - \lambda^2/2}\, d\gamma &= \frac{\lambda^2}{2}\Lambda(\lambda) \\ &\geq \int e^{\lambda F - \lambda^2/2}(\lambda F - \lambda^2/2)\, d\gamma - \Lambda(\lambda)\log\Lambda(\lambda) \\ \Lambda(\lambda)\log\Lambda(\lambda) &\geq \int e^{\lambda F - \lambda^2/2}(\lambda F - \lambda^2)\, d\gamma = \lambda\Lambda'(\lambda). \end{aligned}$$

We define the Herbst argument $H(\lambda) = \frac{1}{\lambda}\log\Lambda(\lambda)$. Then

$$\lambda^2\Lambda(\lambda)H'(\lambda) = \lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) \leq 0.$$

Since $\lambda^2\Lambda(\lambda) > 0$ for $\lambda > 0$, we have $H'(\lambda) \leq 0$.

We have $H(0) = \frac{\Lambda'(0)}{\Lambda(0)} = \int F\, d\gamma = 0$. Thus $H(\lambda) \leq 0$ for all $\lambda \geq 0$. This gives $\Lambda(\lambda) \leq 1$, and so

$$\int e^{\lambda F}\, d\gamma \leq e^{\lambda^2 L^2/2}.$$

Markov's inequality with $\lambda = r$ gives

$$\mathbb{P}(F(Z) \geq r) \leq e^{\lambda^2/2 - \lambda r} = e^{-r^2/2}.$$

13.4 Talagrand's information-transportation inequality

The quadratic Wasserstein distance between two probability measures μ and ν (on the same space) is

$$W_2^2(\mu,\nu) = \inf_{P_{XY}: P_X=\mu, P_Y=\nu} \mathbb{E}|X-Y|^2$$

We can think of this as a distance between measures, motivated by moving probability mass from one to the other. It is actually a metric (satisfies triangle inequality, etc.). Also, it admits a nice dimension decomposition.

$$\begin{aligned} W_2^2(\mu,\nu) &= \inf \mathbb{E}|X-Y|^2 \\ &\geq \sum_{i=1}^n \mathbb{E}|X_i - Y_i|^2 \\ &= \sum_{i=1}^n W_2^2(\mu_i,\nu_i). \end{aligned}$$

Theorem 13.6 (Talagrand's inequality)**.**

$$2D(\mu\|\gamma) \geq W_2^2(\mu,\gamma).$$

Note that both sides grow "linearly" in dimension n. Contrast this with Pinsker's inequality, where the total variation is bounded by 1 regardless of dimension, rendering it rather unhelpful.

Let $P_{X^n} := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical distribution of $X_1, \ldots, X_N \sim \mathcal{N}(0, I)$. The following are true.

1. $\mathbb{E}W_2(P_{X^n},\gamma) \to 0$.
2. $E_t = \{\mu : W_2(\mu,\gamma) > t\}$ is open in the topology of weak convergence.
3. $g_n : (x_1,\ldots,x_n) \mapsto W_2(P_{x^n},\gamma)$ is $n^{-1/2}$-Lipschitz.

Concentration of Lipschitz functions implies

$$P(W_2(P_{X^n},\gamma) > t) \leq e^{-n(t-\mathbb{E}W_2(P_{X^n},\gamma))^2/2}.$$

Sanov's theorem implies

$$\begin{aligned} -\inf_{\mu\in E_t} D(\mu\|\gamma) &\leq \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(W_2(P_{X^n},\gamma) > t) \\ &\leq -\limsup_{n\to\infty} (t - \mathbb{E}W_2(P_{X^n},\gamma))^2/2 \\ &= -t^2/2. \end{aligned}$$

If $W_2(\mu,\gamma) > t$ (i.e. $\mu \in E_t$) then

$$2D(\mu\|\gamma) \geq t^2.$$

Taking $t = W_2(\mu,\gamma) - \epsilon$ and $\epsilon \to 0$ proves the theorem.

Combining with the previous results gives the nice chain

$$I(\mu\|\gamma) \geq 2D(\mu\|\gamma) \geq W_2^2(\mu,\gamma).$$

13.5 The blowing-up phenomenon

For $B \subset \mathbb{R}^n$, let $B_t := \{x : d(x,B) \leq t\}$ be the t-blowup of B.

Theorem 13.7. Let $B \subset \mathbb{R}^n$. If $t \geq \sqrt{-2\log\gamma(B)}$ where γ is the standard Gaussian measure. Then,

$$1 - \gamma(B_t) \leq \exp\left(-\frac{1}{2}(t - \sqrt{-2\log\gamma(B)})^2\right).$$

Roughly, if B contains a sufficient amount of the mass of γ, then B_t contains almost all of the mass!

Concretely, if $\gamma(B) = 10^{-6}$, then $\gamma(B_{13}) \geq 1 - 3\times 10^{-13}$. If we consider $\mathbb{R}^n$ for n large, most Gaussian vectors lie on a spherical shell of radius $\sqrt{n}$. Note that 13 is a very small distance compared to this $\sqrt{n}$.

References

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