Lecture Notes

STAT 157 with Alexander Strang
Spring 2024

Contents

Acknowledgements I would like to recognize and thank Professor Alexander Strang for all information contained in these notes.

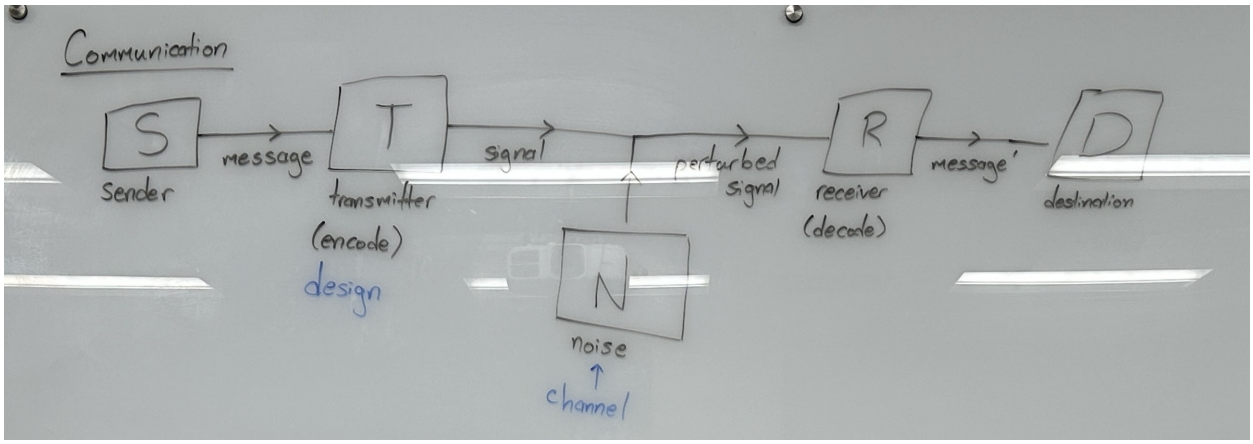
Thanks to Leon Ma who typeset'd the notes for Lecture 9.

Disclaimer: These notes are high-level and mainly capture verbal notes not on the blackboard – there's much more detail on the board with verbal details and drawings. As such, this represents only a subset of what the class covers.

1 Tuesday, January 16th

1.1 Communication

We work with the following model:



which consists of the following entities:

- Sender: you
- Message: What you want to send
- Signal: What you actually sent
- Transmitter: a function that encodes messages to signals
  - This is a design question: we want to make it so that the perturbed signal can be decoded with minimal error
- Source of Noise: different depending on how the channel works, but fixed to the type of channel
- Perturbed signal: signal + noise
- Receiver: they decode the perturbed signal to get message' which is then sent to the destination

1.2 Channel Capacity

Definition: Channel Capacity

max over possible encodings of the rate at which we can send information (messages) with low error.
What is the rate that you can send signals?

Statistical Information: Does not care about content of the message — instead: How many encodings can be represented?
This is as opposed to being length-agnostic where panther has more info than animal.

Problem: observe $Y = y$, estimate the original X given $Y = y$ related by conditional distribution of $X|Y = y$

- Prior: $P(X = x)$
- Posterior: $P(X = x|Y = y)$
- Likelihood: $P(Y = y|X = x)$

Posterior $\propto$ Likelihood $\times$ prior, where alpha means proportional to (there's some factor in the denominator that we are not caring about).

Information: The reduction of uncertainty regarding what was sent given what was received.

- Regarding an unknown after an observation

But now we need to think how do we measure uncertainty in X, $H(X)$?
We say $U[X] = U(p)$ is a functional:

- $f \to I[f]$ is equivalent to $\int f$

Then we can think about the uncertainty in X after an observation, $H(X|Y = y)$.
Finally we could average over all possible received messages to define:

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

We want a function that is a measure of uncertainty that is symmetric/invariant over all possible choices of labels (i.e. random variables).

1.3 Extrinsic vs Intrinsic Measures

Extrinsic measures: Depends on a realization of probability space (i.e. variance). You are measurable.

- Depends on how you label your outcomes.
  - This is useful if you lost your keys in 1 of 3 places and want to know the name of the place where you lost your keys
- Traditional Statistics
- Functions of, expectation of random variables.

Intrinsic measures: Information Theory is the best example of this with uncertainty.

- Depends only on the probability space, the set of outcomes, list of probabilities. Example: $p = [p_1, p_2, \ldots, p_\Omega]$
- $U(p)$ is permutation invariant. That is, $U[p] = U[p']$.
  - This is good if you want to know $P[\text{err}|b]$ and don't want the answer to be dependent on 'a' being before 'b'
  - Also makes sure the same statement in English vs French vs etc. gives the same entropy.

1.3.1 When does Intrinsic matter more than extrinsic?

- Suppose that X is a random variable with finitely many discrete & distinct outcomes, which takes on integers ≥ -1.
- Also suppose we can get exact answers to questions of the form: is $X > \alpha$ for any α.
- How many answers do I need on average to know the value of X.

You can scale the labels (dilate the graph out horizontally) which would change extrinsic measures but not intrinsic measures. $\text{Var}(aX) \neq \text{Var}(X)$ but our Intrinsic question above does not care.

2 Thursday, January 18th

2.1 Basic Definitions and Concepts

Let $\Omega = \{\text{set of all positive outcomes}\}$, $w_j \in \Omega$ where w_j is a randomly chosen outcome.

Let $|\Omega|$ = number of possible outcomes.

Let $X(w_j) = x_j$, $X : \Omega \to \mathcal{X}$ (our representation).

Let $\Pr(w = w_j) = \Pr(X = x) = p_j$.

Definition: Intrinsic Functional

An intrinsic functional is defined as $U[X] = U(p)$, where $U : \Delta_{|\Omega|} \to \mathbb{R}$
Where Δ_n is the simplex which is $\{p \in \mathbb{R}^n, p_i \geq 0 \forall i \text{ and } \sum_{I=1}^{n} p_i = 1\}$ which geometrically looks like this image:

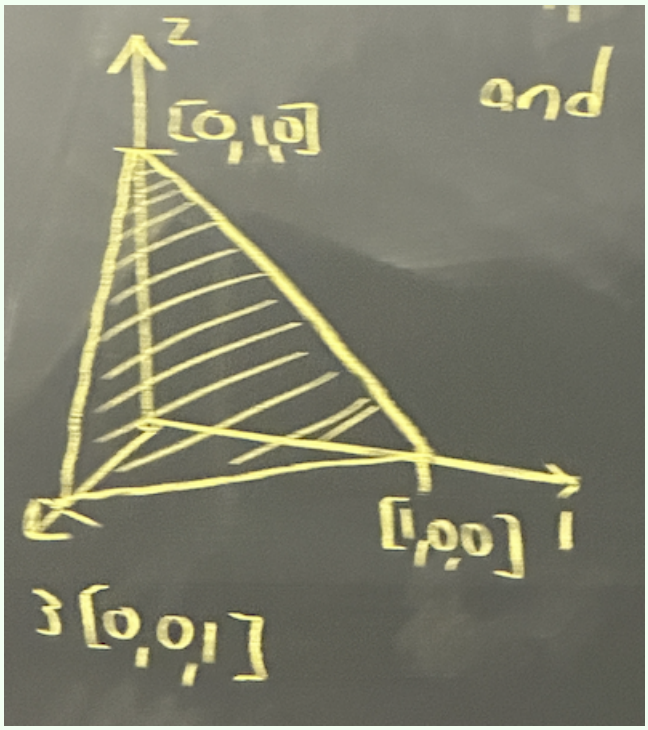


Figure 1: Here we see a 3D-Simplex: a tetrahedron

2.2 Shannon Axioms

Definition: Shannon Axioms

Shannon axioms for a functional $U : \Delta_n \to \mathbb{R}$, intrinsic:

- Regularity/smoothness: $U(p)$ is continuous in p.
  - Allows us to take limits: $\lim_x f(x) = f(\lim_x x)$.
- If $p = [\frac{1}{n}, \ldots, \frac{1}{n}]$ Uniform$[n]$, $|\Omega| = n$, then $U(p)$ is non-increasing in n.
- Uncertainty in X = Uncertainty in Y + Expected Uncertainty in X given Y, for any choice $\{S_1, S_2, \ldots\}$.

2.3 Khinchin Axioms

Definition: Khinchin Axioms

Khinchin axioms for a functional $U : \Delta_n \to \mathbb{R}$, intrinsic:

- U is intrinsic.
- For a given $|\Omega|$, $U(p)$ is maximized when $p = [\frac{1}{\Omega}, \ldots, \frac{1}{\Omega}]$ aka uniform.
- Adding impossible outcomes ($p_j = 0$) shouldn't change your measure of uncertainty $U(p)$.
- Chain Rule: $U[X, Y] = U[Y] + \mathbb{E}_Y(U[X|Y])$
  - Sometimes $\mathbb{E}_Y(U[X|Y])$ is written (mistakenly) as $U[X|Y]$.

2.4 Berkeley's List of Axioms

Definition: Berkeley's Axioms

Berkeley's list of axioms:

- $U(p) = 0$ if $p = [0, 0, \ldots, 0, 1, 0, \ldots, 0]$

2.5 Composition

Composition: If we partition Ω into mutually exclusive and collectively exhaustive subsets $S_1, \ldots, S_\psi$. Then let X represent the outcome in some subset. Let Y be an indicator for the set containing X: $Y = j$ if $X \in S_j$ This is backwards from last class: here we will observe Y and estimate X in 2 stages:

1. Observe Y, we find that X is contained in some say S_1, which tells us that we are contained in a certain subset — but there's still some uncertainty. So now we:
2. Observe X given $Y = y$.

We want $U[X] = U[Y] + \mathbb{E}_y(U[X|Y = y])$.

2.6 Shannon's Entropy Theorem

Definition: Shannon's Entropy Theorem

Given either Shannon's or Khinchin's axioms then:

$$U[X] = U(p) = H[X] = H(p) = \text{“Shannon's Entropy''}$$

$$= -\sum_{\text{all } x \in \mathcal{X}} \Pr(X = x) \log_a(\Pr(X = x))$$

$$= -\sum_{j=1}^{|\Omega|} p_j \log_a(p_j)$$

$$= \mathbb{E}_X[\log_a(1/\Pr(X = x))]$$

Where:

- By convention $p = 0 \implies p\log(p) = 0$ which is known as continuity and is proved by L'Hôpital's as $p \to 0^+$ addressing Khinchin Axiom 3: $H[p] = H[[p, 0, 0, 0]]$ as impossible events have 0 uncertainty.
- a is an arbitrary base which is a choice of unit to measure units (it is not uniquely specified by the axioms).
  - $\log_2 \implies$ “bits”
  - $\log_3 \implies$ “ternary function”
  - $\ln \implies$ “nats”
  - All of these are off by a scaling factor, as shown below,

2.7 Logarithmic Change of Base for Entropy

$$\log_b a = \frac{\log_d a}{\log_d b} \implies H_b(p) = \log_b(a) H_a(p).$$

3 Tuesday, January 23rd

Here is the text converted into valid LaTeX, assuming that the ‘important‘ environment and ‘md-framed‘ package are already defined in your LaTeX document preamble:

“‘latex

3.1 Shannon Entropy

We will start with the definition of Shannon Entropy.

We say an uncertainty functional $U[X]$ satisfying either Shannon or Khinchin's axioms must be a Shannon Entropy, $U[x] = H_a[X] = -\sum_{\text{all } x \in \mathcal{X}} p(x) \log_a(p(x)) = \mathbb{E}_X[\log_a(1/p(x))]$ where a is the choice of units.

3.1.1 Properties of Entropy

- It is intrinsic: Entropy will be symmetric under a certain reflection.
- Entropy is bounded below: $H[X] \geq 0$ with equality if and only if $\exists x$ such that $p(x) = 1$.
- The discrete Entropy is bounded above: $H[X] \leq \log_a(|X|)$ with equality if and only if X is distributed uniformly with $H[X] = \log_a(n) = 1$.
- $H(p)$ is continuous in p (also differentiable in p if we throw out impossible events).
  - So that our uncertainty doesn't jump around so that we can have convergence.
  - You can think about all of your outcomes as groups of the smallest denominators.
- $H(p)$ is a concave function of p.
- Chain Rule: $H[X, Y] = H[Y] + H[X|Y] = H[X] + H[Y|X]$
  - Uncertainty in the joint = Uncertainty in the marginal + Uncertainty in the conditional.
  - We only observe Y first + $\mathbb{E}$ [uncertainty left over] where the latter term is the definition of conditional entropy.

3.1.2 Other Entropies

- **Definition**: Joint Entropy
  Given X, Y with distribution $p(.,.)$, then
  $H[X, Y] = \mathbb{E}_{X,Y}[\log_a(1/p(X, Y))] = -\sum_{\text{all } x, \text{all } y} \log_a(p(x, y))$.
- **Definition**: Conditional Entropy given an observation.
  Given X, Y with distribution $p(.,.)$, then
  $H[X|Y = y] = \mathbb{E}_{X|Y=y}[\log_a(1/p(X|Y = y))]$.
- **Definition**: Conditional Entropy
  Given X, Y with distribution $p(.,.)$, then

$$H[X|Y] = \mathbb{E}_Y[H[X|Y = y]] = \mathbb{E}_{X|Y}[\log_a(1/p(X|Y))]$$

.

- **Alternate**:

$$\begin{aligned} U[X] = \mathbb{E}_X[1 - p(X)] & \\ &= 1 - \mathbb{E}_X[p(X)] \text{ by linearity} \\ &= 1 - \sum_x p(x)p(x) \\ &= 1 - \sum_x p(x)^2 \\ &= 1 - \Pr(X_1 = X_2) \\ &= \Pr(X_1 \neq X_2) \end{aligned}$$

  for i.i.d. $X_1, X_2 \sim p$.

Now we will consider what will happen if we relax the chain rule:

$$\begin{aligned} X \text{ independent of } Y &\implies H[X|Y] = H[X] \\ &\implies H[Y|X] = H[Y] \\ &\implies H[X,Y] = H[X] + H[Y] \end{aligned}$$

We will call the three equations above (*).

Fact we will prove in the future:

$$H[X] = \text{Var}(X) \text{ for } X \sim \mathcal{N}(\mu, \sigma)$$

Theorem: If we replace the chain rule with equation (*) in Khinchin's axioms, then

$$U[X] = H_a^{(\alpha)}[X]$$

Where:

$$H_a^{(\alpha)}[X] = \frac{1}{1-\alpha} \log_a \left(\sum_x p(x)^\alpha \right) = \frac{\alpha}{1-\alpha} \log_a \left(||[p_1, \ldots, p_{|X|}]||_\alpha \right)$$

is known as Rényi Entropy.

If we take $\alpha \to 0$ then we get Hartley Entropy: $H_a^{(0)}[X] = \log_a(|X|)$.

If we take $\alpha \to 1$ then we get Shannon Entropy.

If we take $\alpha \to 2$ then we get the Coincidence/Collision Entropy: $H_a^{(2)}[X] = \log_a \left(\frac{1}{\sum_x p(x)^2} \right) = \log_a \left(\frac{1}{\Pr(X_1 = X_2)} \right)$.

If we take $\alpha \to n \geq 1$ then we get:

$$H_a^{(n)}[X] = \frac{1}{n-1} \log_a \left(\frac{1}{\Pr(X_1 = X_2 = \ldots = X_n)} \right)$$

.

If we take $\alpha \to \infty$ then we get:

$$H_a^{(\infty)}[X] = \log_a \left(\frac{1}{\max_{x \in X} p(x)} \right)$$

.

3.1.3 Information

Information: The reduction of uncertainty after an observation. The mutual information between X and Y is defined as

$$I(X;Y) = \mathbb{E}_Y[H[X] - H[X|Y=y]] = H[X] - \mathbb{E}_Y[H[X|Y=y]] = H[X] - H[X|Y]$$

I am asking about X when I am observing Y.

The information Y carries about X is the uncertainty in X before observation minus the expected uncertainty in X after observation.

3.1.4 Properties of Information

1. Self-Information:
   - **Lemma**: $I(X;X) = H[X]$. This follows from the definition of information with the fact that $H[X|X=x] = 0$ which implies $H[X|X] = 0$ since there is zero expected surprise after observing itself.
   - If we identify an outcome then we've learned everything there is to know about it. The amount of information equals the original uncertainty.
2. Independent: $I(X;Y) =$ prior - posterior $= H[X] - H[X|Y] = H[X] - H[X] = 0$.
3. $I(X;Y) = I(Y;X)$, the information X carries about Y is the same as the information Y carries about X.
   - Proof: $I(X;Y) = H[X] - H[X|Y] = H[X] - (H[X,Y] - H[Y]) = H[X] + H[Y] - H[X,Y] = I(Y;X)$.

4 Thursday, January 25th

Information Theory

Last time we defined information:

$$I(X; Y = y) = H[X] - H[X|Y = y]$$

and mutual information as the expected information before making the observation:

$$I(X; Y) = \mathbb{E}_y[I(X; Y = y)].$$

We called this mutual information because of the third property covered last lecture.

4.1 Relative Entropy

Relative Entropy, also known as KL Divergence, is defined as:

$$D(p||q) = \mathbb{E}_{x\sim p}[\log(p(x)/q(x))].$$

4.2 Mutual Information is a KL Divergence

$$I(X; Y) = I(Y; X) = D(p_{X,Y}||p_X p_Y) = \mathbb{E}_{X,Y}[\log(p(X, Y)/(p(X)p(Y)))].$$

Here, $p_X p_Y$ is the product distribution where $p(X = x)p(Y = y)$.

4.3 Convex Functions

A function is convex if it lies beneath any chord of the function. For a convex combination (meaning $p_1 + p_2 = 1$) $[p_1, p_2] \in \Delta_2$ which is the simplex in 2 dimensions, we have:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2).$$

And the expected value of $f(X)$ is:

$$\mathbb{E}_{X\sim[x_1,x_2] \text{ with probability } [p_1,p_2]}[f(X)] \geq f(\mathbb{E}[X]).$$

We have strict convexity if we have a strict inequality.

4.4 Jensen's Inequality

Generalizing to any dimension, we get Jensen's inequality: If f is convex, for any distribution p,

$$\mathbb{E}_{X\sim p}[f(X)] \geq f(\mathbb{E}_{X\sim p}[X]).$$

4.4.1 Proof via Induction

We can prove Jensen's inequality via induction:

- Base Case: $n = 2$, see above.
- Inductive Step: Assume Jensen's holds for $n = k$.
- Inductive Conclusion: Show that Jensen's for $n = k$ implies Jensen's for $n = k + 1$.

$$\begin{aligned}
\mathbb{E}_{X\sim p \text{ which is } k+1 \text{ outcomes}}[f(X)] &= \sum_{j=1}^{k+1} p_j f(x_j) \\
&= \sum_{j=1}^{k} p_j f(x_j) + p_{k+1} f(x_{k+1}) \\
&= \left(\sum_{m=1}^{k} p_m\right) \sum_{j=1}^{k} \left[\frac{p_j}{\sum_{m=1}^{k} p_m} f(x_j)\right] + p_{k+1} f(x_{k+1}) \\
&\geq (1 - p_{k+1}) f\left(\mathbb{E}_{X \neq X_{k+1}}[X]\right) + p_{k+1} f(x_{k+1}) \\
&= f\left(\sum_{j=1}^{k+1} p_j x_j\right) = f(\mathbb{E}[X]). \text{ QED.}
\end{aligned}$$

4.5 Information Inequality

The information inequality states that $I(X;Y) \geq 0$. The proof is as follows:

$$\begin{aligned}
I(X;Y) &= D(p_{X,Y} || p_X p_Y) \\
&= \mathbb{E}_{X\sim p}[\log(p(x)/q(x))] \\
&= \mathbb{E}_{X\sim p}[-\log(q(x)/p(x))] \\
&= \mathbb{E}_Y[-\log(Y)] \quad \text{where} \quad Y(X) = \frac{q(x)}{p(x)} \text{ is convex, so by Jensen's Inequality,} \\
&\geq -\log(\mathbb{E}_{X\sim p}[q(x)/p(x)]) \\
&= -\log(\sum_x p(x) \cdot \frac{q(x)}{p(x)}) \\
&= -\log(\sum_x q(x)) \\
&= -\log(1) = 0. \quad \text{QED.}
\end{aligned}$$

5 Tuesday, January 30th

5.1 End of Unit 1: Three Inequalities

We start by ending unit 1 with three inequalities:

1. $H[X] + H[Y] \geq H[X, Y]$
   - Interpretation: Joint entropy of independent X, Y is greater than or equal to the joint entropy of X, Y.
   - Equality requires: Independence between X and Y.
   - Proof: $H[X] + H[Y] \geq H[X, Y] \implies H[X] + H[Y] \geq H[Y] + H[X|Y] \implies H[X|Y] \leq H[X] \implies 0 \leq H[X] - H[X|Y] \implies I(X;Y) \geq 0$, proved yesterday.
2. $H[X|Y] \leq H[X]$
   - Interpretation: Observation reduces uncertainty.
   - Equality requires: Independence between X and Y.
   - Proof: $H[X|Y] \leq H[X] \implies 0 \leq H[X] - H[X|Y] \implies I(X;Y) \geq 0$, proved yesterday.
3. $H[X|Y] + H[Y|X] \leq H[X, Y]$
   - Interpretation: Joint entropy of X, Y is greater than or equal to the expected uncertainty in one given the other.
   - Equality requires: Independence between X and Y.
   - Proof: $H[X|Y] + H[Y|X] \leq H[X] + H[Y|X] = H[X, Y]$, where we first use 2, then chain rule.

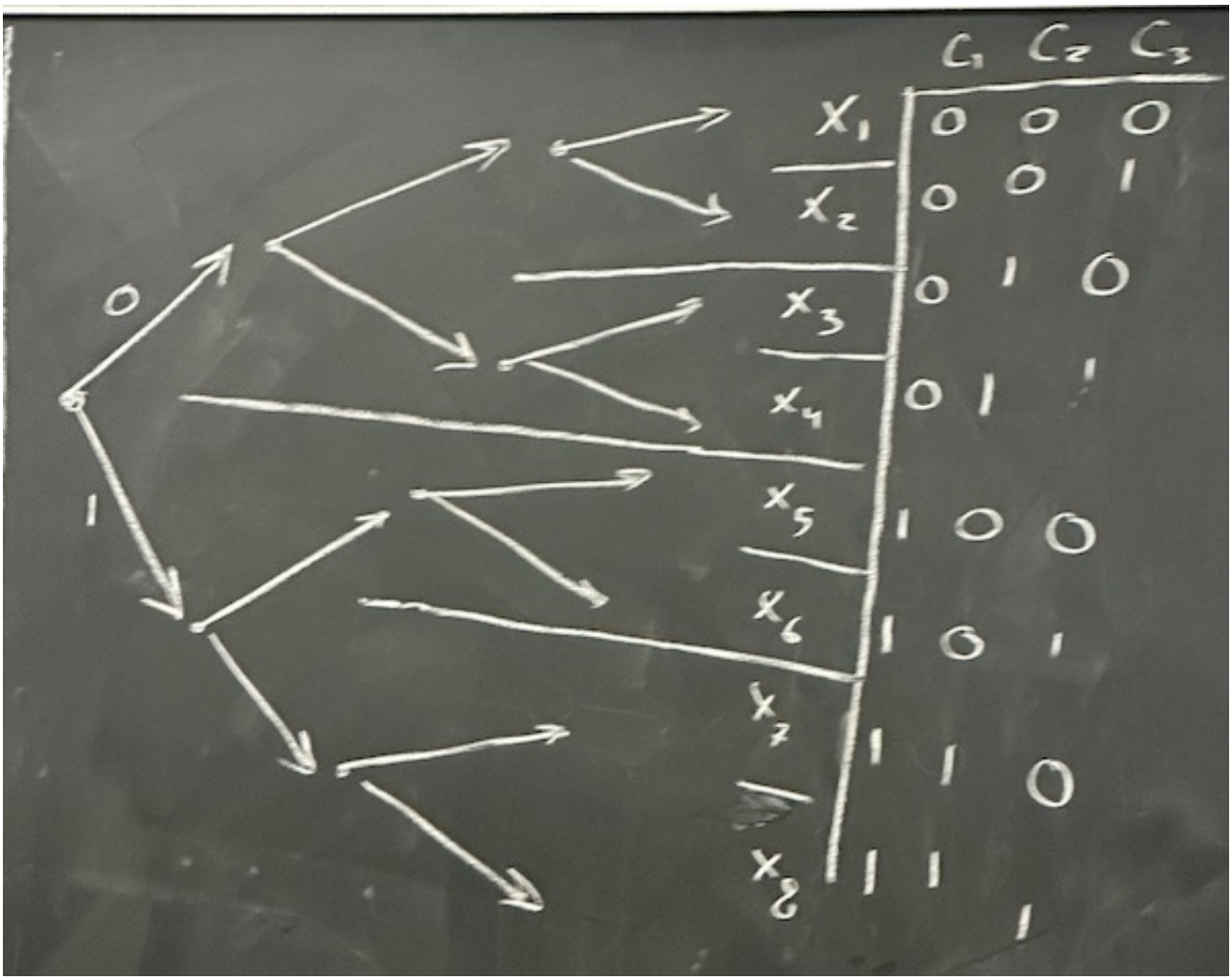
5.2 Motivating Question: Why the log?

To answer this, we will play a game of 20 questions:

1. What are the most efficient questions to ask?
2. How efficient is it?

5.3 Now we play a game:

- Q1: Is it alive? Yes, which we denote as $Y_1 = 1$.
- Q2: Is it an animal? Yes, which we denote as $Y_2 = 1$.
- Q3: Is the word 2 syllables? Yes, which we denote as $Y_3 = 1$.
- Q4: Is it a Narwhal? No, which we denote as $Y_4 = 0$.
- Q5: Does it live in the ocean? No, which we denote as $Y_5 = 0$.
- Q6: Is it an egret? No, which we denote as $Y_6 = 0$.
- (Answer: It was a rabbit).



5.4 Motivating Question: Random Number Generation

Given an RNG:

1. How to efficiently process/transform randomness?
2. How efficient?

If we have $|\mathcal{X}| = n$ outcomes, then we represent $d^m = n$, which requires a length $l = \lceil \log_d(n) \rceil$.

Suppose $|\mathcal{X}| = n$, $p \in \Delta_n$, $p = [p_1, p_2, \ldots, p_n]$, $p_j \geq 0$, $\sum_{j=1}^m p_j = 1$, a sequence of approximations $\{\tilde{p}^{(k)}\}^\infty$, if $\lim_{k\to\infty} \tilde{p}^{(k)} \to p$, then, $U(\tilde{p}^{(k)}) \to U(p)$ as $k \to \infty$.

Restrict $\tilde{p}^{(k)}$ to be rational, as they are dense in reals. This allows us (per the Chain Rule) to group microscopic events into macroscopic ones.

There's an equivalence between search trees and coding. In either case, you naturally get a logarithmic measure (from the depth of the tree).

6 Thursday, March 1st

6.1 The Emergence of the Logarithm in Information Theory

Today we will show that the logarithm appears due to the smoothness/continuity axiom as well as the chain rule axiom. Given the assumption of continuity, we know that the (microscopic event model — from thermodynamic statistical mechanics). If $p = [5/8, 1/4, 1/8]$.

6.2 Optimizing Codings and Tree Representations

Then we will look at optimizing codings/tree representations of encoding and bounds on the best you can do:

- Kraft's Inequality
- Entropy as efficiency bound
- Towards optimality…

6.3 The Chain Rule and Its Implications

If groups of Y filter to a unique X, $Y \sim \text{Unif}[1, M]$, $X = x(Y)$. Then $U[X, Y] = U[Y] = U[X] + U[Y|X]$. But we have asymmetry:

$$U[X] = u(M) - \sum_j P(X = x_j) \cdot u(m_j)$$

thus $U[\cdot]$ isn't specified by how u behaves on equally likely outcomes.

We define $\{u(m)\} = U[\text{uniform dist. over } m \text{ outcomes}]$. Suppose $m = kl$, $m, k, l \in \mathbb{Z}$. e.g., $m = 10$, $k = 5$, $l = 2$. Then we can partition the top row of $m = 10$ events into $l = 2$ sets which each contain $k = 5$ things. Chain rule: $u(m) = u(kl) = U[X] + U[Y|X]$

$$= u(l) + \mathbb{E}_x[U(Y|X = x)]$$

$$= u(l) + \mathbb{E}_x[u(k)] = u(k) + u(l)$$

.

$$u(m) = u(kl) = u(k) + u(l)$$

.

$u = \log$ is then forced upon us as logs turn multiplication into addition.

Or more rigorously, suppose $m = D^l$ for some $D, l \in \mathbb{Z}$, e.g., $m = 8$ equally likely things. This can be done as 4 sets, each of size 2, which using the chain rule gives us that $u(D^{l-1}) + u(D)$. We can then recurse, each time lowering the largest exponent (which was originally l) by 1. This means the depth of our tree will be l.

$$u(m) = u(D^l) = u(D \cdot D^{l-1}) = u(D^{l-1}) + u(D) = u(D^{l-2}) + u(D) + u(D) = \ldots = l \cdot u(D) + u(1)$$

and we know that $u(1) = 0$ per Khinchin's Axiom. Now we have $u(D^l) = l \cdot u(D)$, which we also could've got without the axiom if we only recurse $l - 1$ times as $u(D^{l-(l-1)}) = u(D)$. Thus $u(D^l) = \log_D(m) \cdot u(D) = m = D^l$, then it must be true that $U[\text{uniform distribution over } |X|]$ is proportional to $\log(|X|)$.

6.4 Coding

A set of $\mathcal{X}$ represents, $X \in \mathcal{X}$, $X \sim p$, $|X| < \infty$ (we extend in the textbook to countably infinite sets). Define: a code $C : \mathcal{X} \mapsto D^k$, where:

1. $D^* = \{\text{set of all possible strings with symbols in a } d\text{-ary alphabet}\}$
2. $c(x) =$ codeword for x, e.g., $c(x) = 0110$ each symbol is specifying a choice in the tree
3. $l(x) = |c(x)| =$ length of codeword = number of symbols used number of questions answered = the length of the path

Define: $L = \mathbb{E}_{X \sim p}[l(X)] \iff$ average number of steps needed.

6.5 Instantaneous Codes (Prefix Codes)

- A mapping $\{x_1, x_2, x_3, x_4, \ldots\} \mapsto C^*$ where $C(\{x_1, x_2, x_3, \ldots\}) = c(x_1)c(x_2)c(x_3)\ldots$
- Non-singular: $x \neq x' \implies c(x) \neq c(x')$
- No prefixes: No codeword is the prefix of another, meaning all codewords are leaves of the tree.
- Example: $c(x_1) = [0, 0, 1]$, $c(x_2) = [0, 0, 1, 1]$

We can draw a binary tree with 0 as up and 1 as down, to represent this coding scheme.

6.6 Kraft's Inequality

Kraft's inequality states that for a finite or countably infinite set of outcomes $\mathcal{X}$ you want to encode, and a D-ary dictionary, then all instantaneous codes must satisfy:

$$\sum_{i=1}^{n} D^{-\ell_i} \leq 1.$$

Conversely, for any given set of natural numbers $\ell_1, \ell_2, \ldots, \ell_n$ that satisfy the above inequality, there exists a uniquely decodable code over an alphabet of size D with those codeword lengths. This applies to all uniquely decodable codes, which include instantaneous codes.

Even better, there's a recipe to build such a code!

6.7 Quiz Time!

This section is left intentionally blank for the quiz content.

7 Tuesday, February 6th

7.1 Kraft's Inequality and Code Construction

Kraft's inequality points to an explicit construction. The action of us being able to actually construct it gives us an application from the inequality. Today we will prove it for $|\mathcal{X}| < \infty$, instantaneous codes, list of lengths $\ell = \{\ell_1, \ell_2, \ldots, \ell_{|\mathcal{X}|}\}$. This is a code without prefixes so you don't have to wait until you get the full code to realize if you have an antecedent or a child. Thus all messages are at leaves.

This is most interesting when all the paths do not have the same lengths. We assign outcomes to some leaves, but not all possible leaves (a fact which is key for the inequality of Kraft), since we didn't say in our encoding that all leaves should be used so we can skip some leaves if the answer to the question is only ever one answer. The usefulness of this is seen more easily in tertiary or higher trees. Now if we denote $\ell_{\max} = \max_x\{\ell(x)\}$.

If we completed the tree then we would have $D^{\ell_{\max}}$ leaves on the $\ell_{\max}$ 'vertical bar' if we draw the tree horizontally.

Question: If we take a stopping node, at depth ℓ, how many children would it have had at depth $\ell_{\max}$?

Answer: $D^{\ell_{\max}-\ell}$.

Expanding on this, if we sum over every x, we get the following inequality: $\sum_x D^{\ell_{\max}-\ell(x)} \leq D^{\ell_{\max}}$. Dividing both sides by $D^{\ell_{\max}}$ gives Kraft's inequality.

7.2 Constructive Approach

Let's go through an example: Given a binary tree with $\ell = \{1, 2, 4, 4\}$, this satisfies Kraft's inequality with the sum being $\frac{7}{8}$ which tells us we are inefficient.

X	c(x)
1	0
2	10
3	1100
4	1101

Claim: If $X \sim p$, $X \in \mathcal{X}$, $|\mathcal{X}| < \infty$.

- Let $L = \mathbb{E}_X[\ell(X)]$.
- Over all $\ell = \{\ell_1, \ell_2, \ldots, \ell_{|\mathcal{X}|}\}$ such that Kraft's inequality is satisfied then, for any code (i.e., any ℓ's satisfy) $L \geq H_D[X]$.

Proof:

$$\begin{aligned} L - H_D[X] &= \mathbb{E}_X[\ell(x) + \log_D(p(X))] \\ &= \mathbb{E}_{X\sim p}[\log_D(p(x)/D^{-\ell(x)})] \quad \text{which resembles a relative entropy} \end{aligned}$$

Define $q(x) = D^{-\ell(x)}/Z$, where $Z = \sum_x D^{-\ell(x)} \leq 1$,

$$\begin{aligned} &= \mathbb{E}_{X\sim p}[\log_D(p(x)/(q(x)\cdot Z))] \\ &= \mathbb{E}_{X\sim p}[\log_D(p(x)/q(x)) - \log_D(Z)] \\ &= D(p||q) - \log_D\left(\sum_x D^{-\ell(x)}\right) \\ &\geq 0 \quad \text{since both } D(p||q) \geq 0 \text{ and } -\log_D\left(\sum_x D^{-\ell(x)}\right) \geq 0. \\ &\implies L \geq H_D[X]. \quad \text{QED.} \end{aligned}$$

We want short codewords to correspond to likely events and vice versa. Mathematically: $D^{-\ell(x)} \approx p(x)$.

7.3 Small Group Activities: Programming & Fano Codes

Now in small groups, we will tackle three activities:

1. Show that the bounds are tight (can be achieved by some scheme) and what is the cost of misspecification? (What is the smallest expected code length using the Shannon code for the wrong prior?)

- We have no inefficiency if $\sum_x D^{-\ell(x)} = \sum_x p(x) = 1 \implies \log_D\left(\sum_x D^{-\ell(x)}\right) = 0 \implies \ell(x) = -\log_D(p(x))$. But ℓ may not be an integer, so we will want to round up so we don't exceed our bound, thus $\ell(x) = \lceil -\log_D(p(x)) \rceil$. If we define $H_D[X] \leq L^* = \min_C\{\mathbb{E}_X[\ell(X)]\} \leq \mathbb{E}_X[\lceil -\log_D(p(x)) \rceil]$

$$< \mathbb{E}_X[-\log_D(p(x))] + 1 \quad \text{since} \quad \lceil -\log_D(p(x)) \rceil < -\log_D(p(x)) + 1$$

$$= H_D[X] + 1.$$

  We can replace the 1 with $\frac{1}{n}$ by amortizing the wasted bit spread over n bits, encoded at once.
- This is known as a Shannon code: assigning short codes to likely events, long codes to unlikely events: $\ell(x) \approx -\log_D(p(x))$.

2. Design an optimal code (Huffman) for English words (see data on bcourses).
   - Let's start by testing the difference between $L = \mathbb{E}_X[\ell(x)]$ and $H_2[X]$, which just means the base 2 entropy.
   - This is known as a Huffman code: Save the last bit/question for the least likely events.
3. Challenge: given $\{X_1, X_2, \ldots\}$, $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p, 1-p)$ design a mapping $f : X \to Y$ (where $Y = f(X)$) such that:
   - $\{Y_1, Y_2, \ldots\}$, $Y_j \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{1}{2}, \frac{1}{2}\right)$
   - Maximizing efficiency: $\eta = \frac{\mathbb{E}[\text{number of output bits, } |Y|]}{|X|}$ where $|X|$ = number of unfair bits
   - That works and is optimal for all $p \in [0, 1]$.
   - This is known as a Fano code: maximize expected information gain per question/bit.

8 Thursday, February 8th

8.1 Logistics

- Refined learning goals due February 15.
- Quiz 2 (on Coding + AEP) next Thursday.

8.2 Three Perspectives on Entropy

We will start with three perspectives on Entropy:

1. Compression: The shortest description.
2. Randomness: A measure of randomness that cannot be created by any deterministic processing.
3. Typically (the AEP): Typical events have $\log(p) \approx$ entropy.

8.3 Efficiency

Mappings: $Y = f(X)$

Efficiency is given by the formula:

$$\eta(f \,|\, |X| = \mu, p) = \eta(f; |X| = \mu, p) = \frac{\mathbb{E}[|Y|]}{m} = \frac{\mathbb{E}[|f(X)|]}{m}$$

Question: What is the maximum of η subject to f?

"Uncertainty in the output $\leq$ uncertainty in the input."

$$\begin{aligned} H[X] &= H[X_1, \ldots, X_m] \\ &= \sum_{j=1}^{m} H[X_j] \quad \text{by independence} \\ &= m \cdot H[X_1] \quad \text{where } H[X_1] = H(p, 1-p) \end{aligned}$$

Suppose Y is uniform over 2^n:

$$
\begin{aligned}
H[Y] &= H[|Y|] + H[Y \,|\, |Y|] \quad \text{by conditioning on } |Y| \\
&= H[|Y|] + \mathbb{E}_n[H[Y \,|\, |Y| = n]] \\
&= H[|Y|] + \mathbb{E}_Y[|Y|] \\
&\leq H[X] \quad \text{Using the inequality proved below.} \\
&= m \cdot H[X_1]
\end{aligned}
$$

This then allows us to say:

$$\eta = \frac{\mathbb{E}[|f(X)|]}{m} \leq H[X_1] - \frac{1}{m} H[|Y|]$$

which asymptotically makes the last term go to 0. Thus $\eta \leq H[X_1]$.

8.4 The Data Processing Inequality

Claim: For any deterministic mapping f, if we let $Y = f(X)$, then the amount of randomness in Y, $H[Y] \leq H[X]$.

Suppose f is injective, which means unique inputs lead to unique outputs, then $H[X] = H[Y]$ for any choice of labels.

Instead, we group inputs for a non-injective f (where f is not one-to-one). Then

$$
\begin{aligned}
H[X] &= H[Y] + H[X \,|\, Y] \quad \text{and since } H[X \,|\, Y] \geq 0, \\
&\quad \text{we know that } H[X] \geq H[Y].
\end{aligned}
$$

8.5 Asymptotic Equipartition Property (AEP)

The Asymptotic Equipartition Property (AEP) states that most sample n-sequences of an ergodic process have probability about 2^{-nH} and that there are about 2^{nH} such typical sequences. If $X_1, \ldots, X_n$ and X_j are iid~p:

$$-\frac{1}{n} \log \left(\underbrace{p(X_1, \ldots, X_n)}_{p(X_1) \cdot \ldots \cdot p(X_n)} \right) \xrightarrow{\text{i.p.}} H[X_1] \quad \text{as} \quad n \to \infty.$$

$$-\frac{1}{n} \sum_{j=1}^{n} \log(p(x_j)) \to -\mathbb{E}_X[\log(p(X))].$$

8.5.1 Typical Set

The typical set $A_\varepsilon^{(n)}$, if $x_1, \ldots, x_n \in A_\varepsilon^{(n)}$ then $x_1, \ldots, x_n$ is typical:

$$A_\varepsilon^{(n)} = \{x_1, \ldots, x_n \mid -\log(p(x_1, \ldots, x_n)) \in [H[X] - \varepsilon, H[X] + \varepsilon]\}.$$

8.5.2 Properties

1. $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H[X]+\varepsilon)} \implies p(x_1, \ldots, x_n) \in [2^{-n(H[X]+\varepsilon)}, 2^{-n(H[X]-\varepsilon)}]$ which can be read as: "elements of the typical set are all $\mathcal{X}$ equiprobable"
2. $\Pr\left[x^{(n)} \in A_\varepsilon^{(n)}\right] \geq 1 - \varepsilon \equiv \Pr\left[x_1, \ldots, x_n \in A_\varepsilon^{(n)}\right] \xrightarrow{n\to\infty} 1$ for any $\varepsilon > 0$ with high probability.
3. $\left|A_\varepsilon^{(n)}\right| \in [(1-\varepsilon)2^{n(H[X]-\varepsilon)}, 2^{n(H[X]+\varepsilon)}]$ for n sufficiently large.

8.5.3 Typicality

If $X_1, \ldots, X_n$ and we let H be the number of heads.

$$p(x_1, \ldots, x_n) = p^{h(x)}(1-p)^{n-h(x)}$$

$$\log(p(x_1, \ldots, x_n)) = \log\left(p^{h(x)}(1-p)^{n-h(x)}\right) = h(x)\log(p) + (n - h(x)) \cdot \log(1-p)$$

Typicality cares about the multiplicity of outcomes.

9 Tuesday, February 13th

9.1 Goals/Main Question: Is Differential Entropy still intrinsic?

- Is Differential Entropy extrinsic?
- Defn.:

Definition: Differential Entropy

Given $X \sim f_X$, then $h[X] = h(f_X)$.

9.1.1 Side-by-Side comparison: Discrete vs Differential Entropy

Discrete:

$$\begin{aligned} H[X] = H(p) &= -\mathbb{E}_X[\log(p(X))] && \text{where } X \sim p \\ &= -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \end{aligned}$$

which we see is intrinsic.

Differential:

$$\begin{aligned} H[X] = h(f_X) &= -\mathbb{E}[\log(p(X))] && \text{where } X \sim f_X \\ &= -\int_{x \in \mathcal{X}} f_X(x) \log(f_X(x)) \, \mathrm{d}x \end{aligned}$$

which we see is extrinsic.

9.1.2 Example with Differential Entropy

Example: $X \sim f_X$, $X \in \mathcal{X} \subseteq \mathbb{R}$

- Let $Y = aX$, then $f_Y(y) = f_X\left(\frac{y}{a}\right) \cdot |a|$
- Then

$$\begin{aligned} h[aX] &= h[Y] \\ &= \mathbb{E}_Y[\log(f_Y(Y))] \\ &= \mathbb{E}_X[\log(f_X(X/a)|a|)] \\ &= \mathbb{E}_X[\log(f_X(X)) + \log(|a|)] \\ &= h[X] + \log(|a|) \end{aligned}$$

9.2 Axiomatic Approach

- Assume: X is continuous and f_X is continuous on $\mathcal{X}$.

- Definition: $X^{(n)} \xrightarrow{n\to\infty} X$ i.i.d. if the density $f_{X^{(n)}} \to f_X$, pointwise almost everywhere (a.e.).

Definition: "Weak convergence"

$\mathbb{E}[g(X^{(n)})] \to \mathbb{E}[g(X)]$ as $n \to \infty$.

9.2.1 Axioms

(i) Axiom 1: We want h to be continuous; we want $h[X^{(n)}] \to h[X]$, if $X^{(n)} \to X$ in distribution.

Defn.:
- $Z^{\Delta X}$ be drawn.
  (a) Draw $Y^{\Delta X}$
  (b) Draw $Z^{\Delta X} \mid Y^{\Delta X} = k$, uniformly from the k^{th} bin.
- Then $Z^{\Delta X} \overset{\text{i.d.}}{\to} X$ as $\Delta X \to 0$.

(ii) Axiom 2 (Uniform/Scale): $h[\text{Uniform on } X] = \log(|X|)$.

(iii) Axiom 3 Chain Rule: Same as the discrete case.

Applying this, we get:

$$h[X] \simeq h[Z^{\Delta X}] = H[Y^{\Delta X}] + \mathbb{E}_y[\ \underbrace{h[Z^{\Delta X}|Y^{\Delta X} = y]}_{\text{uniform bin width } \Delta X(y)}\ ] \qquad [\text{where } \Delta X \to 0,\ Z^{\Delta X} \to X.]$$
$$= H[Y^{\Delta X}] + \mathbb{E}_{Y^{\Delta X}}[\log\{\Delta X(Y^{\Delta X})\}]$$

10 Thursday, February 15th

10.1 Entropy in Different Contexts

Given $X \sim f_x$ then $h[X] = f(f_x)$.

10.1.1 Discrete vs. Continuous

- Discrete: distribution p (list of probabilities).
- Continuous: density f_x.

In the latter case, it is not enumerable. To address this, we note that both are defined as a measure, and we will explore how to define a 'better' measure.

10.1.2 Measures and Calculus

Measures are functions which map sets to probabilities. These sets, subsets of Ω, are called "events". For calculus, we will require a differential dx which is extrinsic here since x is extrinsic.

10.1.3 Entropy Definitions

- Discrete: $H[X] = H(p) = -\mathbb{E}_X[\log(p)] = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$. This is intrinsic.
- Differential: $h[X] = h(f_x) = -\mathbb{E}_X[\log(f_X(x))] = -\int_{x \in \mathcal{X}} f_X(x) \log(f_X(x))\, dx$. This is extrinsic.

- We would need an infinite number of questions to get the answer here.
- We can approach the problem by playing the game of 20 Questions to get within a sufficient region of the answer, which we define as precision.
- Starting with a wider interval means more questions are needed.

The main difference between the two is that after a monotone transformation T, they will be intrinsic and extrinsic respectively.

10.1.4 Density and Precision

$$\text{density}(x) = \frac{\text{Prob. that } X \text{ near to } x}{\delta x} = \frac{\text{Prob.}}{\text{unit length}} \text{ per.}$$

Note: $\Delta x(y)$ denotes the differential in x, for the bin that contains y.

10.2 Quiz Time!

This section is left intentionally blank for the quiz content.

11 Tuesday, February 20th

11.1 Reading and Quiz Announcements

- Reading: Shannon 3, finish T&C Ch. 2.
- Quiz 3 this Thursday on differential entropy.
- Project Part 3 released tonight.

11.2 Goals

Finish up Differential Entropy.

11.3 Last Class Recap

Last class we left off with the following argument: We tried to make three axioms corresponding to the discrete case:

1. Continuity: h is continuous. $z^{\Delta x}$ i.i.d. converges to f_X as $\Delta x \to 0$.
2. Extrinsic: If $X \sim \text{Unif}[\mathcal{X}]$, then $h[X] = \log(|\mathcal{X}|)$.
3. Chain Rule.

Thus, $h[X]$ as $\Delta x \to 0$ approximately equals $h[z^{\Delta x}]$ which equals $H[Y^{\Delta x}] + \mathbb{E}_y[h[z^{\Delta x}|Y^{\Delta x} = y]]$ and equals $H[Y^{\Delta x}] + \mathbb{E}_{Y^{\Delta x}}[\log(\Delta x(Y))]$, which by Riemann integration, is approximately $-\int_{\mathcal{X}} f_X(x)\log(f_X(x))\,dx$ and equals $-\mathbb{E}_X[\log(f_X(x))]$ (the expected surprise).

Where the uncertainty in which bin contains x equals the number of bins, and the uncertainty in $z^{\Delta x}$ given the bin approximates the uncertainty left over in X given $Y^{\Delta x}$, as we refine to higher precision/resolution.

We are matching relative entropies in the limit...

11.4 Properties of Differential Entropy

Properties of Differential Entropy that are not shared by discrete entropy:

1. Unbounded below: For $X \sim p$, $X \in \mathcal{X}$, and $|\mathcal{X}|$, let $p = \text{Uni}(|S|)$, then $h_d[X] = \log_d(|S|) \to -\infty$ as $|S| \to 0$.
2. Extrinsic: We have a reference ground.

Consider transformations $T : Y = T(X)$, for invertible T.

- Example 1: $T(x) = ax + b$, then $h[Y] = h[X] + \log(|a|)$.

- Example 2: For $x \in \mathbb{R}^n$, $T(X) = Ax + b$, $A \in \mathbb{R}^{n \times n}$, invertible, then we get the change of density formula.

11.5 Gaussian Distribution in Signal Processing

Now we look at one of the most common distributions in Signal Processing, Real-world applications, etc: $X \sim \mathcal{N}(\mu, \Sigma)$, what is $h[X]$? $X = T(y)$, $y \sim \mathcal{N}(0, I)$ where $X = AY + b$ for positive definite Σ such that Σ^{-1} and hence A^{-1} exists. Let $b = \mu$, $\Sigma = AA^T$ as $\text{Cov}[X] = \text{Cov}[AY]$. If we realize that $\det(\Sigma^{0.5}) = \det(\Sigma)^{0.5}$ and:

$$h[Y] = -\mathbb{E}_Y[\log(f_Y(y))] = -\mathbb{E}_Y[\log((2\pi)^{-n/2}\exp(-\frac{1}{2}\sum_{j=1}^{n} y_j^2))] = -\mathbb{E}_Y[-\frac{n}{2}\log(2\pi) - \underbrace{E[Y_j^2]}_{1}] = \frac{n}{2}\log(2\pi e).$$

Then we can simplify:

$$h[X] = \frac{1}{2}\log(\det(\Sigma)) + \frac{n}{2}\log(2\pi e).$$

where

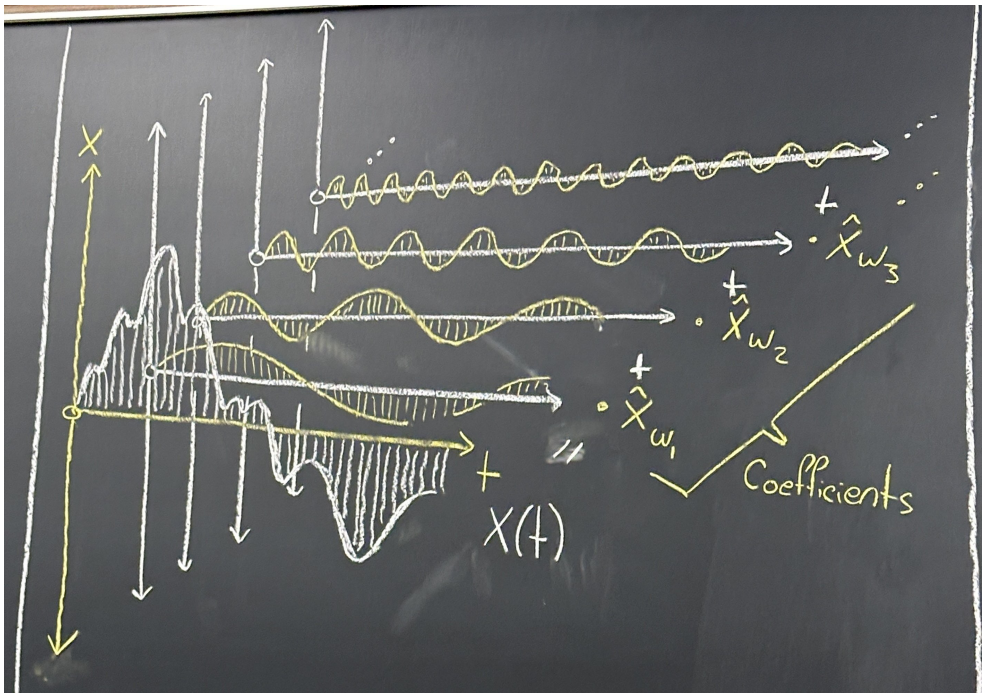
$$h[X] = 1/2 \quad \underbrace{\log(\det(\Sigma))}_{\text{Cov scaling extrinsic dimension}} + \underbrace{n/2}_{\text{dimension}} \underbrace{\log(2\pi e)}_{\text{Gaussian}}.$$

12 Thursday, February 22nd

12.1 Time-Frequency Duality

Draw $\{\hat{X}_j\} \sim f_X, \quad \hat{X} \in \hat{\mathcal{X}}$

Imagine you have a set of basis functions, $\{b_j(t)\}_{j=1}^{n\to\infty}$,
then draw a set of coeff.'s $\{\hat{X}_j\}_{j=1}^{n\to\infty}$.
Set: $X(t) = \sum_{j=1}^{n} \hat{X}_j b_j(t)$.



! Fourier transforms lose temporal resolution.

12.2 Stochastic Processes

Definition: Stochastic Process

is a random function $X(t)$.

Usually you have a SDE (Stochastic differential equation) or a Markov process which is a random operation run conditionally fwd in time.

Here we are generating it all at once by generating a set of coeff.'s. – this gives some curve thoroughout space.

12.2.1 Gaussian Processes

If we draw the set of coeff.'s to be MVG: draw $\hat{X} \sim \mathcal{N}$.

Definition: Gaussian Process

$X(t)$ is a Gaussian Process if:
for any set of samples $\{t_j\}_{j=1}^n$
the r.v. $\vec{X} = \begin{bmatrix} X(t_1), X(t_2), \ldots, X(t_n) \end{bmatrix}$, $X_j = X(t_j)$ is drawn jointly from a MVG.

It is a reasonable class of functions to be working with since it is:

1. Tractable
2. Moderately General. Note that it is not completely general, but it is one of the most simple models that gives us this level of generality.

A MVG is uniquely specified by its mean vector ($\mu(t) = \mathbb{E}[X(t)]$) and its covariance matrix

$$\Sigma = K(t,s) = \text{Cov}(X(t), X(s)) = K(s,t) \text{ s.t. } X \sim \mathcal{N}\left(\begin{bmatrix} \mu(t_1) \\ \mu(t_2) \\ \vdots \\ \mu(t_n) \end{bmatrix}, \begin{bmatrix} K(t_1,t_1) & K(t_2,t_2) & \cdots \\ K(t_2,t_1) & K(t_2,t_2) & \\ \vdots & & \ddots \end{bmatrix} \right)$$

$$K(t,s) = \text{Cov}[X(t), X(s)] = \sum_{j,j=1}^{n} E\left[(\hat{x}_i - \hat{\mu}_i)(\hat{x}_j - \hat{\mu}_j) \right] b_i(t) b_j(s) = \sum_{ij=1}^{n} \left[\hat{\sum} \right]_j b_i(t) b_j(s)$$

mean $\mu(t) = \mathbb{E}[X(t)]$

$$= [b_1(t), \ldots b_n(t)]\, \hat{\varepsilon} \begin{bmatrix} b(s) \\ \vdots \\ b_n(s) \end{bmatrix} = \vec{b}(t)^\top \hat{E} \vec{b}(s)$$

Suppose $\hat{x}_i \perp\!\!\!\perp \hat{x}, \forall i \neq j$:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1^2 & & & \\ & \hat{\sigma}_2^2 & & \\ & & \ddots & \\ & & & \hat{\sigma}_n^2 \end{bmatrix}, \text{ then: } K(t,s) = \sum_{j=1}^{n} \hat{\sigma}^2, b_j(t) b_j(s).$$

12.3 Mercer's Theorem

For most reasonably defined $K, \exists$ an expansion of the kind above for some, $\{b\}_{j=1}^{n \to \infty}$.

12.4 Bochner's Theorem

If $X(t) \sim X(t+s)$ for any s then b_j's are Fourier Modes.

12.5 Shannon-Nyquist Theorem

If $X(t)$ does not contain any frequencies higher than some band limit B, then $X(t)$ can be fully recovered (all the information in $X(t)$ are captured) by finitely many samples sufficiently close, $\Delta t < \frac{1}{2B}$.

12.6 Entropy/Information:

1. In terms of samples,
2. In terms of the coeff. $\hat{X}$

$\mathcal{F}^{-1} : 1 \mapsto 2$ and $\mathcal{F} : 2 \mapsto 1$. Furthermore Shannon gives us this $\mathcal{F}$.

Question: How do I know that I can sample from fractional timesteps (finer samples).

This is a common question in compression, assume you are trying to store a song

Answer: Naively we could just store everything.

Well, $\vec{X} = \begin{bmatrix} X(t_1), \ldots, X(t_n) \end{bmatrix}$.
$\vec{X} \sim \mathcal{N}(\vec{\mu}, K)$

$h[\vec{X}] = \frac{1}{2}\log((2\pi e)^n |K|) = \frac{1}{2}\log(|\det(K)|) + \frac{n}{2}\log(2\pi e)$
Suppose:

$$
\begin{aligned}
I(\vec{X}, \vec{X}) &= h(\vec{X}) - h(\vec{X}' \mid \vec{X}) \\
&= h(\vec{X}) + h(\vec{X}) - h(\vec{X}', \vec{X}) && \text{[Chain Rule]} \\
&= \frac{1}{2}\left[\log(|K_{X'X'}|) + \log(|K_{XX}|) - \log(\left|\begin{bmatrix} K_{X'X'} & K_{X'X} \\ K_{XX'} & K_{XX} \end{bmatrix}\right|)\right] \\
&= -\frac{1}{2}\log\left(1 - \rho^2\right)
\end{aligned}
$$

Smoothness of underlying GP related to MI on refined sampling.

- $X \sim \mathcal{N}(\mu, \Sigma)$, 2D.
- If Σ has SVD: $\Sigma = U \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} U^\top$.

13 Tuesday, February 27th

13.1 Is Information Intrinsic?

Is Information Intrinsic?

- Discrete entropy is intrinsic
- Differential entropy is extrinsic

- Differences in entropy are scale-independent, but extrinsic

- Information is the expected difference in entropy before and after an observation... is information intrinsic?

- recall:
  1. Scaling: if $Y = AX + b$ then $h[Y] = h[X] + \log(|\det(A)|)$ for any invertible A
  2. Δh ⫫ b scaling: if $Y = AX + b$, $Y' = AX' + b$, $X \sim p$, $X' \sim p'$ then:
     $$h[Y] - h[Y'] = h[X] - h[X'] + \underbrace{\log(|\det(A)|) - \log(|\det(A)|)}_{\text{cancel!}}$$
     $$= h[X] - h[X']$$
     does not depend on A or b
  3. generic transform: $Y = T(X)$ for T invertible, differentiable (T is a diffeomorphism)
     - let $\frac{d}{dx} T(x)$ = Jacobian of T at x
     - then (change of density): $f_Y(y) = f_X(x)\,|\det(\frac{d}{dx} T(x))|^{-1}$ at $x = T^{-1}(y)$ $(y = T(x))$
     - so:
     $$h[Y] = -\mathbb{E}_Y\left[\log(f_Y(Y))\right] = -\mathbb{E}_X\left[\log(f_X(x)\,|\det(\dots)|^{-1})\right] = -\mathbb{E}_X\left[\log(f_X(x))\right] - \mathbb{E}_X\left[\log(|\det(\dots)|^{-1})\right]$$
     $$= h[X] - \mathbb{E}_X\left[\log(|\det(\dots)|^{-1})\right] = h[X] + \mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx} T(x))|)\right]$$
     - that is:
     $$h[Y] = h[T(X)] = h[X] + \underbrace{\mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx} T(x))|)\right]}_{\text{expected expansion factor}}$$
  4. Δh is extrinsic: $Y = T(X)$, $Y' = T(X')$, $X \sim p$, $X' \sim p'$

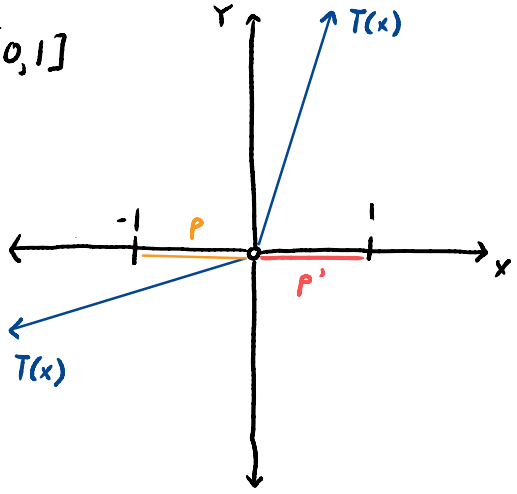
then: $$h[Y] - h[Y'] = h[X] - h[X'] + \underbrace{\mathbb{E}_{X\sim p}\left[\log(|\det(\tfrac{d}{dx} T(X))|)\right] - \mathbb{E}_{X\sim p'}\left[\log(|\det(\tfrac{d}{dx} T(X'))|)\right]}_{\text{average expansion factor over } p \text{ and } p'}$$

$p \neq p'$ so averages may differ...

- Ex: $\mathcal{X} = [-1, 1]$, $p = \text{Uniform}[-1, 0]$, $p' = \text{Uniform}[0, 1]$

$$T(x) = \begin{cases} 3x & \text{if } x > 0 \\ \frac{1}{3}x & \text{if } x < 0 \end{cases}$$

$$\frac{d}{dx} T(x) = \begin{cases} 3 & \text{if } x > 0 \\ \frac{1}{3} & \text{if } x < 0 \end{cases}$$



so $\mathbb{E}_{x \sim p}\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right] = \log(1/3)$

$\mathbb{E}_{x' \sim p'}[\quad \ldots \quad] = \log(3)$

} difference is $\log(1/3) - \log(3) = -\log(9) \neq 0$.

- Information is a difference in entropies so it must be scale independent, is it extrinsic or intrinsic?

- Consider: $Y = T(X)$, $X \sim p_X$

observe z, $X \mid z = z \sim p_{X \mid z = z}$

Then: $I[Y; z] = h[Y] - h[Y \mid z]$

$= h[Y] - \mathbb{E}_z\left[h[Y \mid z = z]\right]$

$= h[X] + \mathbb{E}_x\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right] - \mathbb{E}_z\left[h[X \mid z = z] + \mathbb{E}_{x \mid z = z}\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right]\right]$

$= \quad \ldots \quad - h[X \mid z] - \mathbb{E}_z\left[\mathbb{E}_{x \mid z = z}\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right]\right]$

$\mathbb{E}_{x,z}$ jointly, $\perp\!\!\!\perp$ of z; $= \mathbb{E}_x$

$= h[X] + \mathbb{E}_x\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right] - h[X \mid z] - \mathbb{E}_x\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right]$

$= I[X; z] + \mathbb{E}_x\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right] - \mathbb{E}_x\left[\log\left(\left|\det\left(\frac{d}{dx} T(x)\right)\right|\right)\right]$

Cancel !!

$= I[X; z]$

so, if T is a diffeomorphism (invertible, differentiable)

then $I[Y;Z] = I[T(X);Z] = I[X;Z]$ ← does not depend on T

therefore, independent of representation

(up to diffeomorphisms)

- **Conclusion:**

  (i) differences in h are scale independent, but extrinsic

  (ii) information is intrinsic ($\$$ scale independent), at least up to diffeomorphism!

- Ex: $I[X;Y]$ is independent of $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$

  (see Gaussian examples)

- Suggests there should be a unifying definition of discrete and differential info...

  - if $x \in X$, $y \in Y$, let P be a partition of X

    Q ... of Y

  - let $X^{(P)}$, $Y^{(Q)}$ be the coarse-grained X $\$$ Y

- **Def:** $I(X;Y) = \sup_{P,Q} \left\{ I(X^{(P)};Y^{(Q)}) \right\}$

  $I(X;Y)$: possibly diff. $\quad$ $I(X^{(P)};Y^{(Q)})$: discrete

  $\sup_{P,Q}$: "max over all partitions"

, unlike h, I is a limit of discrete approximations

and, coarse-graining never increases information shared by X and Y

(usually discards some info)

13.2 Gaussian Examples: Mutual Information for Random Vector

(b) dimension on the outside...

$$h[X] = n \ln\left(\det\left(\sqrt{2\pi e}\left(\Sigma^{1/2}\right)^{1/n}\right)\right) = n \ln\left(\sqrt{2\pi e}\,\det\left(\Sigma^{1/2}\right)^{1/n}\right)$$
$$\approx n \ln\left(4\,\det\left(\Sigma^{1/2}\right)^{1/n}\right)$$

(since $\ln(x) = \frac{n}{n}\ln(x) = n\ln(x^{1/n})$)

- Interpretation: $\det(\Sigma^{1/2}) = \prod_{j=1}^{n} \sigma_j$ so $\det(\Sigma^{1/2})^{1/n} = \left(\prod_{j=1}^{n}\sigma_j\right)^{1/n} = \bar{\sigma}$

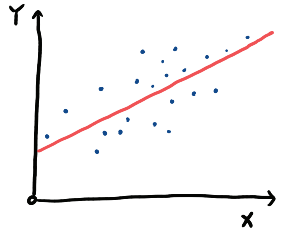
where $\bar{\sigma}$ = geometric average of the principal s.d.'s

Then: $h[X] = n \ln\left(\sqrt{2\pi e}\,\bar{\sigma}\right) \approx n \ln(4\bar{\sigma})$ (scale; dimension; standard normal)

- $X \sim N(\mu, \Sigma)$ is as uncertain as
  - (i) n i.i.d. draws from uniform on interval length $4\bar{\sigma}$
  - (ii) n i.i.d. draws from normal w/ s.d. $\bar{\sigma}$

Computing Mutual Information - Example II: Correlated Gaussian Variables

- Aim: Compute $I[X;Y]$ for:
  1. X & $Y \sim N$ univariate
  2. X & $Y \sim N$ multivariate



1. Univariate: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu, \sigma_Y^2)$, $\mathrm{Corr}[X,Y] = r_{xy}$

- What is $I[X;Y]$?

$$I[X;Y] = \begin{cases} h[X] - h[X|Y] \\ h[Y] - h[Y|X] \\ h[X] + h[Y] - h[X,Y] \end{cases}$$

($X|Y=y$ is N $\forall\, y$; $Y|X=x$ is N $\forall\, x$) — could compute; would need formula for conditionals

← only need formula for entropy of N's

$$I[X;Y] = h[X] + h[Y] - h[X,Y]$$

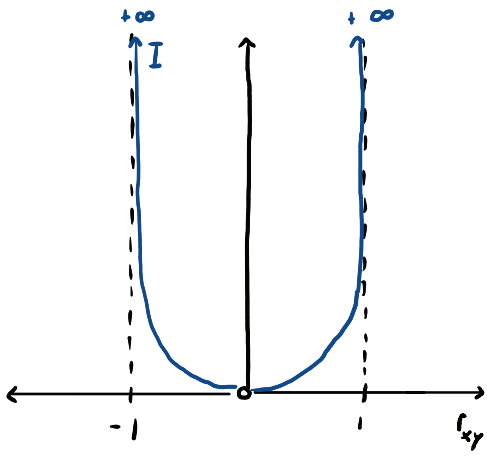
$$= \tfrac{1}{2}\ln(2\pi e\,\sigma_x^2) + \tfrac{1}{2}\ln(2\pi e\,\sigma_y^2) - \tfrac{1}{2}\ln\left((2\pi e)^2 \det\left(\begin{bmatrix}\sigma_x^2 & \sigma_x\sigma_y r_{xy}\\ \sigma_x\sigma_y r_{xy} & \sigma_y^2\end{bmatrix}\right)\right)$$

Σ for $\begin{bmatrix}X\\Y\end{bmatrix}$

$$= \left(\tfrac{1}{2}\ln(2\pi e) + \tfrac{1}{2}\ln(2\pi e) - \tfrac{2}{2}\ln(2\pi e)\right) + \tfrac{1}{2}\ln(\sigma_x^2) + \tfrac{1}{2}\ln(\sigma_y^2) - \tfrac{1}{2}\ln\left(\sigma_x^2\sigma_y^2 - \sigma_x^2\sigma_y^2 r_{xy}^2\right)$$

$$= \ln(\sigma_x) + \ln(\sigma_y) - \ln\left(\sigma_x\sigma_y\sqrt{1-r_{xy}^2}\right)$$

$$= \ln\left(\frac{\sigma_x\sigma_y}{\sigma_x\sigma_y\sqrt{1-r_{xy}^2}}\right) = -\tfrac{1}{2}\ln(1-r_{xy}^2)$$



so: $I[X;Y] = -\frac{1}{2}\ln(1-r_{xy}^2)$ if $X \sim N$, $Y \sim N$ (univariate) and $\mathrm{Corr}[X,Y] = r_{xy}$

the more correlated
the larger the mutual information

- notice:
  - (i) if $r_{xy} = 0$, $X \perp\!\!\!\perp Y$, $I[X;Y] = 0$
  - (ii) if $|r_{xy}| \to 1$, Y predicts X exactly, $I \to \infty$
  - (iii) I is independent of μ_x, μ_y and σ_x, σ_y } intrinsic? or just scale invariant?
    ... scale invariant, depends only on correlation...

2. Multivariate: $X \sim N$ n-dimensional, $Y \sim N$ m-dimensional

$$\begin{bmatrix}X\\Y\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}\right)$$

where $\Sigma = \left[\begin{array}{c|c}\Sigma_{XX} & \Sigma_{XY}\\\hline \Sigma_{YX} & \Sigma_{YY}\end{array}\right]$ (n rows, m rows; n columns, m columns)

$$\Sigma_{XX} = \mathrm{Cov}[X], \quad \Sigma_{YY} = \mathrm{Cov}[Y]$$
$$\Sigma_{XY} = \mathrm{Cov}[X,Y] = \Sigma_{YX}^T.$$

then: $I[X;Y] = h[X] + h[Y] - h[X;Y]$

$$= \tfrac{1}{2}\left(\ln(\det(\Sigma_{XX})) + \ln(\det(\Sigma_{YY})) - \ln(\det(\Sigma))\right)$$

can we reduce?

- let $Var[X] = [\mathbb{V}[X_1], \mathbb{V}[X_2], \dots \mathbb{V}[X_n]]$

  $Var[Y] = [\mathbb{V}[Y_1], \mathbb{V}[Y_2], \dots \mathbb{V}[Y_m]]$

  ↑ "variance of"

- let $D_v = diag(v_1, v_2, \dots)$

- then: $\Sigma_{XX} = D_{var[X]}^{1/2} R_{XX} D_{var[X]}^{1/2} = \begin{bmatrix} SD[X_1] & & \\ & SD[X_2] & \\ & & \ddots \\ & & & SD[X_n] \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{bmatrix} \begin{bmatrix} SD[X_1] & & \\ & SD[X_2] & \\ & & \ddots \\ & & & SD[X_n] \end{bmatrix}$

  $r_{ij} = Corr[X_i, X_j]$

  $\Sigma_{YY} = \underbrace{D_{var[Y]}^{1/2}}_{SD} \underbrace{R_{YY}}_{Corr} \underbrace{D_{var[Y]}^{1/2}}_{SD}$

and: $\Sigma = \begin{bmatrix} D_{var[X]}^{1/2} & \\ & D_{var[Y]}^{1/2} \end{bmatrix} \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{XX} \end{bmatrix} \begin{bmatrix} D_{var[X]}^{1/2} & \\ & D_{var[Y]}^{1/2} \end{bmatrix}$

↑ $R_{YX_{ij}} = Corr[Y_i, X_j]$

- now: $\log(\det(\Sigma)) = \log(\det(D_v^{1/2})^2 \det(R)) = \log(\det(D_v)) + \log(\det(R))$

so: $\log(\det(\Sigma_{XX})) + \log(\det(\Sigma_{YY})) - \log(\det(\Sigma))$

$= [\log(\det(D_{var[X]})) + \log(\det(D_{var[Y]})) - \log(\det([\begin{smallmatrix} D_{var[X]} & \\ & D_{var[Y]} \end{smallmatrix}]))$

$+ \log(\det(R_{XX})) + \log(\det(R_{YY})) - \log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}]))]$

↳ $= \log(\prod_{i=1}^{n} \mathbb{V}[X_i] \prod_{j=1}^{m} \mathbb{V}[Y_j]) - \log(\prod_{i=1}^{n} \mathbb{V}[X_i] \prod_{j=1}^{m} \mathbb{V}[Y_j]) = 0$ ... all var. terms cancel out

$= \log(\det(R_{XX}) \det(R_{YY})) - \log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}]))$

so: $I[X;Y] = \frac{1}{2}[\log(\det(R_{XX}) \det(R_{YY})) - \log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}]))]$

$= \frac{1}{2}[\log(\det([\begin{smallmatrix} R_{XX} & 0 \\ 0 & R_{YY} \end{smallmatrix}])) - \log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}]))]$

$= -\frac{1}{2}[\log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}])) - \log(\det([\begin{smallmatrix} R_{XX} & 0 \\ 0 & R_{YY} \end{smallmatrix}]))]$

$= -\frac{1}{2}[\log(\det([\begin{smallmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{smallmatrix}])) + \log(\det([\begin{smallmatrix} R_{XX}^{-1} & 0 \\ 0 & R_{YY}^{-1} \end{smallmatrix}]))]$

so: $I[X;Y] = \frac{1}{2}\left[\log\left(\det(R_{XX})\det(R_{YY})\right) - \log\left(\det\left(\begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{bmatrix}\right)\right)\right]$

$= \frac{1}{2}\left[\log\left(\det\left(\begin{bmatrix} R_{XX} & 0 \\ 0 & R_{YY}\end{bmatrix}\right)\right) - \log\left(\det\left(\begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{bmatrix}\right)\right)\right]$

$= -\frac{1}{2}\left[\log\left(\det\left(\begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{bmatrix}\right)\right) - \log\left(\det\left(\begin{bmatrix} R_{XX} & 0 \\ 0 & R_{YY}\end{bmatrix}\right)\right)\right]$

$= -\frac{1}{2}\left[\log(\ \ldots\ ) + \log\left(\det\left(\begin{bmatrix} R_{XX}^{-1} & 0 \\ 0 & R_{YY}^{-1}\end{bmatrix}\right)\right)\right]$

$= -\frac{1}{2}\left[\log(\ \ldots\ ) + \log\left(\det\left(\begin{bmatrix} R_{XX}^{-1/2} & 0 \\ 0 & R_{YY}^{-1/2}\end{bmatrix}\right)^2\right)\right]$

$= -\frac{1}{2}\log\left(\det\begin{bmatrix} R_{XX}^{-1/2} & 0 \\ 0 & R_{YY}^{-1/2}\end{bmatrix}\det\begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{bmatrix}\det\begin{bmatrix} R_{XX}^{-1/2} & 0 \\ 0 & R_{YY}^{-1/2}\end{bmatrix}\right)$

$= -\frac{1}{2}\log\left(\det\left(\begin{bmatrix} R_{XX}^{-1/2} & 0 \\ 0 & R_{YY}^{-1/2}\end{bmatrix}\begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{bmatrix}\begin{bmatrix} R_{XX}^{-1/2} & 0 \\ 0 & R_{YY}^{-1/2}\end{bmatrix}\right)\right)$

$I[X;Y] = -\frac{1}{2}\log\left(\det\left(\begin{bmatrix} I_{n\times n} & M \\ M^T & I_{m\times m}\end{bmatrix}\right)\right)$ where $M = R_{XX}^{-1/2} R_{XY} R_{YY}^{-1/2}$

- so $I[X;Y]$ is independent of all X & Y variances (scale independent), and depends "M"

to simplify, use the block determinant formula

$$\det\left(\begin{bmatrix} A & B \\ C & D\end{bmatrix}\right) = \det(A)\det(D - CA^{-1}B) \quad \text{if } A \text{ \& } D \text{ are square...}$$

so:

$$\det\left(\begin{bmatrix} I & M \\ M^T & I\end{bmatrix}\right) = \det(I)\det(I - M^T I M) = \det(I - M^T M) = \prod_{j=1}^{m}(1 - \sigma_j(M)^2)$$

($\sigma_j(M)$: singular values of M)

so: given $X, Y \sim N$ w/ correlations R_{XX}, R_{XY}, R_{YY}

$$I[X;Y] = -\frac{1}{2}\log\left(\det(I - M^T M)\right) = -\frac{1}{2}\log\left(\prod_{j=1}^{m}(1 - \sigma_j(M)^2)\right) \quad \text{where } M = R_{XX}^{-1/2} R_{XY} R_{YY}$$

- fully generalizes the 1D case

$$I[X;Y] = -\frac{1}{2}\log(1 - r_{xy}^2)$$

- in both cases I only depends on the correlations, is scale independent.

14 Thursday, February 29th

14.1 Information Inequalities II: Data Processing and Fano

Information Inequalities: (commentary on 3 key inequalities)

- Def: $I[X;Y] = D(P_{X,Y} \| P_X P_Y)$

- exploit Jensen, log-sum, & convexity of D to show (proofs in T3C)...

1. Information is Nonnegative: $I[X;Y] \geq 0$ and $= 0$ iff X and Y are independent
   (on average, observation reduces uncertainty...
   $H[X|Y] \leq H[X]$, $H[X,Y] \leq H[X] + H[Y]$)
   * we've seen this before...

2. Data-Processing Inequality:

- Def: $X \to Y \to Z$ form a (3-step) Markov chain if Z is conditionally independent of X given Y
  i.e. $Pr(Z=z|Y=y, X=x) = Pr(Z=z|Y=y)$
  more generally: $X(t_1) \to X(t_2) \to \dots \to X(t_n)$ is a discrete-time Markov chain if $X(t_{j+k})$ is conditionally independent of $X(t_{j-s})$ given $X(t_j)$ $\forall\ k>0,\ s>0$
  (future ⫫ of past given present)

- Can use as a model for all data processing... receive Y, perform a deterministic or random procedure on Y that does not use any other info about X (Y is all our data regarding X), output Z
  Z represents processed Y

- Inequality: If $X \to Y \to Z$ is a Markov chain, then:
  $I[X;Z] \leq I[X;Y]$
  so, if $Z = g(Y)$, $I[X;T(Y)] \leq I[X;Y]$ for any g.
  (transform performed by processing)
  (processing never ↑ information.)
  * more general case of the concavity of I... (see T3C 2.7.4)
  w/ equality iff X is conditionally independent of Y given Z

We saw this in the deterministic processing case for entropy... see coin tossing conversion example.
let $Y=X$, $Z = T(Y)$, $I[X;T(X)] \leq I[X;X] = H[X]$
(T: some transform)

3. Fano's Inequality: let $X \sim p$, $X \in \mathcal{X}$. Observe Y. Use Y to estimate X via an estimator $\hat{X} = g(Y)$ for some g, possibly stochastic.

- Inequality: $Pr(\hat{X} = g(Y) \neq X) \geq \dfrac{H[X] - I[X;Y]^{*} - 1}{\log(|\mathcal{X}|)}$
  ($Pr(\hat{X} = g(Y) \neq X)$: probability of an error in estimation)
  * $H[X] - I[X;Y] = H[X|Y]$. information reduces conditional uncertainty after observing Y, reduces estimation error.

so,

1. Information is nonnegative and only zero if independent
2. Processing never increases information
3. Information reduces the probability of estimation error

1. and 3. are not too surprising.

2. (data-processing) is strikingly strong and somewhat confounding since we usually process data
   - indeed, almost all of stats & data science is about processing data to learn from it...

} so, let's dig into 2. a bit...

(the following section is ~ philosophy/argument, my aim here is to help answer the question: why is information intrinsic?)

- Data-Processing Interrogated:

  - let's start with an idea the data-processing inequality clarifies...

    - Def: Suppose $X \sim P_\theta$ where P_θ depends on an unknown parameter θ then:

      $g(X)$ is a sufficient statistic if X is conditionally independent of θ given $g(X)$.

    - then: (i) $\theta \to X \to g(X)$ is a Markov chain for any g (by construction)

      so, for most statistics g, $I[\theta; g(X)] \leq I[\theta; X]$

      (we lose info about the unknown by processing X to produce $g(X)$, unless...

      (ii) if $g(X)$ is sufficient:

      $\theta \to g(X) \to X$ is a Markov chain (X cond. $\perp\!\!\!\perp$ of θ given $g(X)$)

      then: $I[\theta; X] \leq I[\theta; g(X)]$ by data-processing

      together: $I[\theta; X] \overset{\text{(ii)}}{\leq} I[\theta; g(X)] \overset{\text{(i)}}{\leq} I[\theta; X]$ ... requires $I[\theta; g(X)] = I[\theta; X]$.

      ($\uparrow$ sufficient; $\uparrow$ statistic (i.e. processed X))

- So, data processing, $X \to g(X)$ loses information about θ unless $g(X)$ is sufficient and, if $g(X)$ is sufficient, processing loses no information...

- $I[\theta; g(X)] = I[\theta; X]$ iff $g(X)$ is sufficient.

  - Ex: $\{X_i\}_{i=1}^n \sim \text{Poisson}(\lambda)$ for unknown λ then $g(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient

    (indeed, the sample mean is sufficient for $\text{Exp}(\lambda)$, $N(\mu, \sigma^2)$ for known σ

    $\{X_i\}_{i=1}^n \sim \text{Gamma}(r, s)$ for r, s unknown then $g(X) = \left[\frac{1}{n}\sum_{i=1}^n X_i, \left(\prod_{i=1}^n X_i\right)^{1/n}\right]$ is sufficient (sample arithmetic and geometric means).

- What if we can't find a sufficient statistic (the usual case)?
  then data-processing loses information!

- Contradicts (?) common intuition that we learn from data (extract information) by processing it

  - Ex: low-dimensional embedding/latent space representation

$$\underset{\text{high-D}}{X} \xrightarrow{\text{process}} \underset{\text{low-D}}{g(X)}$$

- usually s.t. $g(X)$ has "intrinsically" meaningful coordinates
- i.e. the position of data, axes, or relative position of data after applying g is meaningful
  (think: PCA, interpretable feature selection, word embedding, etc.)

- Notice: processing is all about changing representation
  information is intrinsic, representation agnostic
  - All of the original information is in X, we just might not know how to read/decode it...
    - We process data to change into a representation we understand...

- Ex: - encryption - all the information is in the encrypted message but need to decrypt (process) to read it

  - translation - you recieve a message in a language you don't read, all the information is there, it just needs to be translated to read

  - feature identification - often we measure many different variables without knowing a priori which are important/ what combination of variables matters again, all the information exploited is in the original data (we can't produce information unless we change our prior model - i.e. bring in outside info), the task is to change representation so that we can interpret the data.

- We process data to "decrypt" the information stored in it
  i.e. change representation so that we can interpret the data

  - this is why we built an intrinsic theory... to distinguish the information stored in data, independent of any encryption (lossless, deterministic, invertible change of representation) from our ability to interpret the information (represation dependent)

- information is only reduced or preserved by processing

since processing either:

- is deterministic and invertible: information is preserved since we can reconstruct X from $g(X)$
- is sufficient: information is preserved since all desired relations w/ Y are encoded in $g(X)$
- is stochastic or noninvertible and is not sufficient: information is lost since relevant aspects of X can't be recovered from $g(X)$,
  - if $X \neq X' \rightarrow g(X) = g(X')$ then can't reconstruct X from g
  - if g is noisy, can't reconstruct X w/ 100% certainty.

- thus, information theory cannot address the interpretability of the representation
  - it can provide bounds on the accuracy/limits of idealized decryption processes

- interpretability is subjective (I can't read Sanskrit, "it's all greek to me", I can't read binary but my computer can)
- information is objective * (really, intrinsic)

- to compare & contrast information and interpretability as separate quantities we'd need a quantitative definition of interpretability

  - an idea: define interpretability as the computational complexity of some interpretation process that aims to decrypt/decode the message encoded in the data

  - Ex: bootstrapping:

    $\{X_i\}_{i=1}^n \sim P$, $s: \mathcal{X}^n \to \mathbb{R}$ is some statistic of interest
    want to find the quantiles of $s(X)$ for uncertainty quantification

    - data: $\{X_i\}_{i=1}^n \sim P$ - desired quantity: sampling dist of $s(X_1, \dots X_n)$

    given $\{X_i\}_{i=1}^n$ we only have one sample of $s(X_1, \dots X_n)$,
    can't make a histogram/quantiles out of one sample

    - data processing: try bootstrapping: $\{X_i\}_{i=1}^n \longrightarrow g(\{X_i\}_{i=1}^n) = \{\{\underbrace{Y_j^{(k)}\}_{j=1}^n}_{\text{boot strapped samples}}\}_{k=1}^m$

    - interpretation: given $\{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m \longrightarrow \{s(Y_1^{(k)}, \dots Y_n^{(k)})\}_{k=1}^m$
      now have many samples, can compute (estimate) quantiles.

- it seems like we gained information by bootstrapping to get "new" data sets...
  (people will say this)

- But, bootstrapping is just a stochastic/randomized computation technique
  - it's just noisy data processing
- we didn't add new information by adding new randomness
  - all the draws $Y_j^{(k)} \overset{iid}{\sim} \underbrace{\text{Uniform}(\{X_j\}_{j=1}^n)}_{\text{"empirical distribution"}}$, i.e. are from the data we already had
  - $\{X_j\}_{j=1}^n$ fully specifies the dist. of bootstrap samples, thus the sampling dist. of $s(Y_1, \dots Y_n)$.
  - all info used was available in $\{X_j\}_{j=1}^n$

- really, the problem is that we don't know how to "decode" a sampling dist. from one sample
- so, process data to create many datasets $\{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m$
  - we know how to "decode" many sample data sets into quantiles. Cost is $\mathcal{O}(m)$. Cheap for most m.
  - Switching $\{X_j\}_{j=1}^n \to \{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m$ doesn't create info, rather it changes the representation of $\{X_j\}_{j=1}^n$ such that we can easily extract the sampling variability of s.

- in fact, the noisy process (bootstrap) loses information
  - ideally, we would enumerate every possible bootstrap sample and compute s for each, very expensive to fully enumerate, for a given n. Then would lose no information, but the cost would be enormous. Direct evaluation of sampling variability of s w.r.t. the empirical distribution is available w/ no info loss, but too computationally expensive to use

[• trade some information loss in processing for cheaper evaluation (easier interpretation)]

^

this tradeoff is key in cryptography...
safe/private codes should change representation while retaining information, but while requiring expensive decoding

14.2 What is Relative Entropy (KL)? Part 1: Review

What is "relative" entropy?

- Definition: given p, q, p and q both supported on $\mathcal{X}$ and $\text{Supp}(q) = \{x \text{ s.t. } q(x) > 0\} \subseteq \text{Supp}(p) = \{x \text{ s.t. } p(x) > 0\}$ the relative entropy:

$$D(p \| q) = \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(X)}{q(X)}\right)\right].$$

  - also called: "Kullback-Leibler Divergence (KL)"

- motivating questions:
  (i) how can we interpret $D(p\|q)$? What does it mean? (separate from a convenient object arising in analysis)
  (ii) in particular, why is the asymmetry $D(p\|q) \neq D(q\|p)$ a feature of the definition rather than a bug?

- What we already Know:

1. Properties:

(i) Nonnegativity: $D(p\|q) \geq 0$ and $= 0$ if and only if $p = q$ (almost surely for $X \sim p$)
  - loosely, $D(p\|q)$ is a "distance" between p & q
    we'll qualify this

(ii) Asymmetry: $D(p\|q) \neq D(q\|p)$
  - apparent from definition. p and q are not exchangeable ... average over p not q
  - true distances (metrics) should be symmetric
  - not a distance, if we treat as a psuedo-metric then this is just a bug
  - why is this substantive/useful?

(iii) Joint Convexity (see T&C): if $\lambda \in [0,1]$ then

$$D(\lambda p + (1-\lambda)p' \| \lambda q + (1-\lambda)q') \leq \lambda D(p\|q) + (1-\lambda) D(p'\|q').$$

2. Example uses (so far):

(i) excess description length (cost of misspecification in compression):

(a) given code C, $\ell(x) = |C(x)|$, $L(C) = \mathbb{E}_{x\sim p}[\ell(x)]$

- if we can only send n characters (d-ary) per unit time, then, using C, the channel capacity (max-average X sent per unit time) is $\propto 1/L(C)$

- wanted to minimize $L(C)$ (maximizes capacity for noiseless channel)

- saw $L(C) = \underbrace{H[X]}_{\text{theoretical lower bound}} + D(p\|q) - \log(z)$ ← inefficiency from mismatch in code design and distribution of X

where $q(x) = \frac{1}{z} D^{-\ell(x)}$, $z = \sum_x D^{-\ell(x)}$.

(b) if we designed the code C using the assumption $X\sim q$ but $X\sim p$ then (ex: Shannon code)

$$L(C) \in H[X] + \underbrace{D(p\|q)}_{\text{excess description length}} + [0, 1]$$

- interpretation: $D(p\|q)$ is the cost of misspecification
  - true for other tasks
  - asymmetric since p is the true dist, generates samples } average over p not q
  - q is the approximation/proposal, does not generate samples }

(ii) differences in entropies ("relative"):

- Ex: $H[X] - \log(|\mathcal{X}|) \overset{*}{=} -D(p\,\|\,\text{Uniform}[\mathcal{X}])$

* recall from limiting construction of differential entropy

$$\left(H[X] - \log(|\mathcal{X}|) = H[X] + \log(|\mathcal{X}|^{-1}) = \mathbb{E}_{x\sim p}\left[-\log(p(x)) + \log(|\mathcal{X}|^{-1})\right] = -\mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x)}{|\mathcal{X}|^{-1}}\right)\right] = -D(p\|\,\text{Uniform}[\mathcal{X}])\right)$$

so: $H[X] = H[\text{Uniform}(X)] - D(p \,\|\, \text{Uniform}(X))$

difference between max possible entropy (uniform) and entropy of p ... information needed to go from Uniform(X) to p

- this algebra does not work for generic q ($D(p\|q) \neq H(q) - H(p)$ generically)
- works when misspecified...
  - Suppose $X \sim p$ but we think $X \sim q$
  - measure our surprise on sampling X as $-\log(q(X))$
  - our expected surprise is:

$$-\mathbb{E}_{X\sim p}[\log(q(X))] \neq H(q) \text{ or } H(p)$$

if we have many samples could estimate
let $\tilde{H}(p\|q)$ be our estimate to $H[X]$
using q for surprise

- Compare: $\tilde{H}(p\|q) = -\mathbb{E}_{X\sim p}[\log(q(X))]$ w/ $H[X] = H(p)$ ... ("cross entropy", usually $H(q,p)$)
  - $\mathbb{E}$ error in estimate: $\tilde{H}(p\|q) - H(p) = \mathbb{E}_X[-\log(q(X)) + \log(p(X))]$

$$= D(p\|q) \geq 0$$

- so, if $X \sim p$, think $X \sim q$, and estimate $H[X]$ using samples then the expected error in our estimate is $D(p\|q)$.
- the error is >0 if $p \neq q$. Always too/overly surprised (overestimate $H[X]$)
- $D(p\|q)$ is our excess surprise (when misspecified)

3. Relation to Mutual Info: $I[X;Y] = D(p_{X,Y} \,\|\, p_X p_Y)$

so:
(a) mutual info is the expected excess description length of joint draws X, Y if we ignored their dependence

(b) mutual info is the excess surprise when observing joint samples X, Y if we thought they were independent

Contextualize I via D not D...

(c) we will see... if X is an unknown w/ prior p, observe Y, posterior $X|Y=y$, then $D(p_X \,\|\, p_{X|Y=y})$ = the information gained about X via observation.

14.3 What is Relative Entropy (KL)? Part 2: Statistical Interpretation

Relative Entropy is a Divergence (not a distance)

- roughly, a divergence $p \| q$ measures how distinguishable q is from p when we draw data from p
  - If $X \sim p$, but I thought $X \sim q$, how hard would it be to tell?
  - How much data would I need to find out?

... easiest to motivate divergences via hypothesis testing ...

- **Problem:** We observe a sequence of samples $\{X_i\}_{i=1}^n \overset{iid}{\sim} p$ where $p = p_0$ or p_1, but we don't know which...

- **hypotheses:** $H_0: p = p_0$, $H_1: p = p_1$

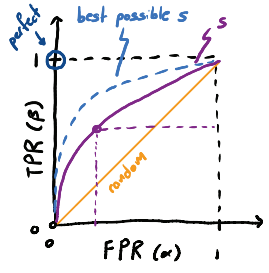
- **Test:**
  1. pick a test statistic $s(x_1, x_2, \dots x_n) : \mathcal{X}^n \to \mathbb{R}$
  2. estimate the sampling distribution of $s(X_1, X_2, \dots X_n)$ under $X \sim p_0$ and $X \sim p_1$
  3. compare observed s to its sampling distribution under p_0 and p_1
  4. define accept (H_1) and reject (H_0) regions using 2. (e.g. a threshold on s for a one-sided test)
  5. reject/accept based on where the observed s is

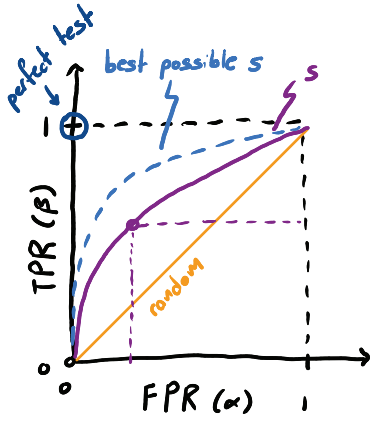
  (e.g. accept if $\Pr(\text{sampling } s(X_1, X_2, \dots X_n) \text{ less likely than observed } s(x_1, x_2, \dots x_n)$ under $H_0) \leq \alpha$ for some α $\Rightarrow$ α controls the FPR... i.e. chance we pick H_1 when H_0 is true i.e. controls the significance of the test)

  could exchange the roles of H_0 and H_1 here, convention is arbitrary...

- Choosing s and the accept/reject region (α) determines the test (fixes decision thresholds on s)
  - thus, also fixes the FNR, i.e. prob. of choosing H_0 when H_1 is true i.e. controls power/sensitivity

- generally want power & significance (sensitivity & specificity)
- for a given s, visualize trade-off w/ ROC
  - plot best TPR (power) possible for a given FPR (significance) equivalently, best TNR for a given FNR





- Want to choose our statistic s to:
  - (·) maximize power (sensitivity) for all significances (specificity) when using H_0 to control (fix accept/reject to control FPR = control significance
  - (··) maximize significance (specificity) for all powers (sensitivity) when using H_1 to control (fix accept/reject to control FNR = control sensitivity

- What s should we choose?

- Neyman-Pearson: if $f: \mathbb{R}^+ \to \mathbb{R}$ is injective (one-to-one) and monotonic and we define:

"likelihood ratio" $LR(x_1, x_2, \dots x_n) = \dfrac{Pr(X_1 = x_1, X_2 = x_2, \dots X_n = x_n \mid H_1)}{Pr(X_1 = x_1, X_2 = x_2, \dots X_n = x_n \mid H_0)}$

if large, data more likely under H_1, choose H_1
if small, ... under H_0 choose H_0

then $s(x_1, x_2, \dots x_n) = f(LR(x_1, x_2, \dots x_n))$ will achieve (·) and (··).

- idea: the best way to choose H_0 vs. H_1 is by comparing the likelihood of the data under each model.

- Ex: (i) LR test (ii) log(LR) test (iii) $H_0: X \sim \exp(\lambda_0)$, $H_1: X \sim \exp(\lambda_1)$ can use $s = \frac{1}{n}\sum_{j=1}^{n} x_j$.

- to understand how distinguishable P_0 and P_1 are, we should study the sampling dist. of $s(X_1 \dots X_n)$
  - at least for large n...

Q: What is $\mathbb{E}[s(X_1, \dots X_n)]$? What is the sampling dist. for large n?

- Single Sample:

$$\mathbb{E}[s(X)] = \begin{cases} \mathbb{E}_{X \sim P_0}[f(LR(X))] & \text{if } H_0 \text{ true} \\ \mathbb{E}_{X \sim P_1}[f(LR(X))] & \text{if } H_1 \text{ true} \end{cases}$$

- in particular: log(LR):

sign flips if use P_0/P_1 for LR

$$\mathbb{E}[s(X)] = \begin{cases} \mathbb{E}_{X \sim P_0}\left[\log\left(\frac{P_1(X)}{P_0(X)}\right)\right] = -D(P_0 \| P_1) & \text{if } H_0 \text{ true} \\ \mathbb{E}_{X \sim P_1}\left[\log\left(\frac{P_1(X)}{P_0(X)}\right)\right] = +D(P_1 \| P_0) & \text{if } H_1 \text{ true} \end{cases}$$

order depends which is true.

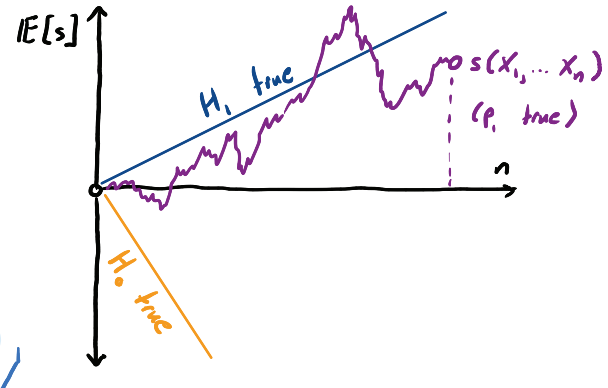
- **interpretation:** $D(p\|q)$ = expected value of log likelihood ratio (best test stat for distinguishing p & q) on a single draw, when p is true

- **multi-sample:** $\mathbb{E}[s(X_1,\dots,X_n)] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[f\left(\frac{P_r(X_1,X_2,\dots X_n\,|\,H_1)}{P_r(X_1,X_2,\dots X_n\,|\,H_0)}\right)\right] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[f\left(\prod_{j=1}^{n}\frac{P_r(X_j\,|\,H_1)}{P_r(X_j\,|\,H_0)}\right)\right]$

  - ideally, choose f s.t. the terms here separate ... log is natural (separate product over each sample)

- **log(LR):** $\mathbb{E}[s(X_1,\dots X_n)] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[\log\left(\prod_{j=1}^{n} LR(X_j)\right)\right] = \sum_{j=1}^{n}\mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[\log(LR(X_j))\right]$

$$= \begin{cases} -n\,D(P_0\|P_1) & \text{if } H_0 \\ n\,D(P_1\|P_0) & \text{if } H_1 \end{cases}$$



- the larger $D(P_1\|P_0)$ the faster we expect to acquire evidence distinguishing P_1 and P_0 when P_1 is true
- ... $D(P_0\|P_1)$ ... when P_0 is true

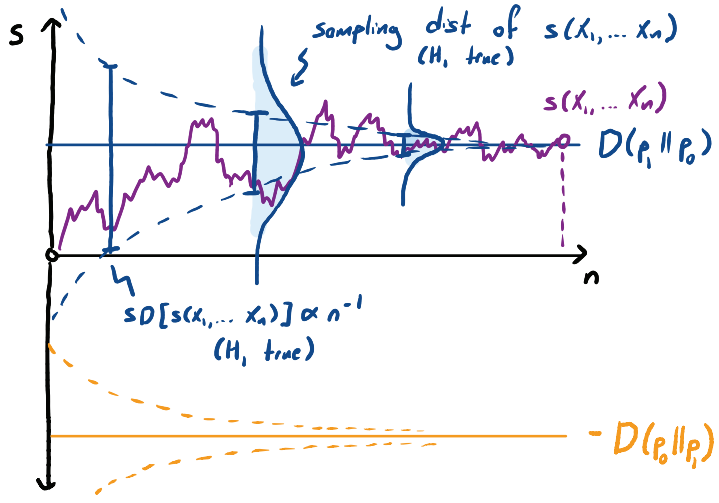
$D(p\|q)$ is the rate we gain evidence to distinguish p & q when p is true

- If we use $f^{(n)}(t) = \frac{1}{n}\log(t) = \log(t^{1/n})$

then:

$$\mathbb{E}[s(X_1,\dots X_n)] = \begin{cases} -D(P_0\|P_1) & \text{if } H_0 \text{ true} \\ +D(P_1\|P_0) & \text{if } H_1 \text{ true} \end{cases}$$

and, $s(X_1,\dots X_n) = \frac{1}{n}\sum_{j=1}^{n}\log(LR(X_j))$ (i.i.d. draws)



so, by the CLT, as $n\to\infty$, the sampling distribution of $s(X_1,\dots X_n)$ converges to a $\mathcal{N}$ distribution w/ $\mathbb{E} = D(\cdot\|\cdot)$ and variance $= \mathcal{O}(n^{-1})$.

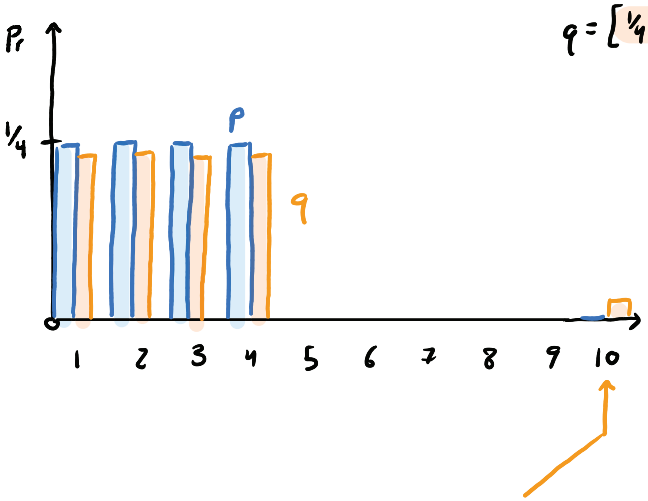
- asymptotically, $D(p\|q)$ and $D(q\|p)$ measure how distinguishable p is from q when drawing from p, or when drawing from q

- **Interpretation**: $D(p\|q) = \mathbb{E}$ of test statistic $(s(x_1,\dots x_n) = \log(LR(x_1,\dots x_n)^{1/n}))$ for the $\log(LR)$ test for distinguishing p & q when p is true.
  - ↕ asymmetric
  - loosely, "distinguishability given data $\sim p$"

- Changing f in $s(x) = f(LR(x))$ induces a family of "divergences" generated by f... (see next section)

- This construction/interpretation explains the asymmetry of D...

- Ex: $\mathcal{X} = \{1, 2, \dots, 10\}$ $p = [1/4, 1/4, 1/4, 1/4, 0, 0, 0, 0, 0, 0]$
  $q = [1/4 \cdot (1-\varepsilon)\ (\times 4), 0, 0, 0, 0, 0, \varepsilon]$
  } only differ by a rare event that is possible under q



rare event: $X=10$, if $X=10$, we know w/ 100% certainty $X \sim q$

we can build examples where
$D(p\|q) \to 0$ but
$D(q\|p) \to \infty$!!!
very asymmetric.

- if p is true:
$$D(p\|q) = \mathbb{E}_{X\sim p}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \sum_{x=1}^{4} \log\left(\frac{1/4}{1/4(1-\varepsilon)}\right)\cdot\frac{1}{4}$$
$$= -\log(1-\varepsilon) \simeq \varepsilon \xrightarrow[\text{as } \varepsilon\to 0]{} 0$$
so, when p is true, it's hard to distinguish q from p

- if q is true,
$$D(q\|p) = \mathbb{E}_{X\sim q}\left[\log\left(\frac{q(x)}{p(x)}\right)\right] = \sum_{x=1}^{4} \log(\dots)\cdot\frac{1}{4}(1-\varepsilon) + \varepsilon \underbrace{\log(\varepsilon/0)}_{\log(\infty)=+\infty}$$
$$= +\infty$$
so, when q is true p & q are easy to distinguish...

- Why? if q is true, we could sample $X=10$ (prob. ε) if we sample $X=10$ we know w/ complete certainty that $X\sim q$ not p since prob $X=10$ if $X\sim p$ is zero!

15 Tuesday, March 5th

15.1 Logistics

Proposal is no longer due this Thursday, now due next Tuesday.

A lot of students are doing the discussion only just before classes. At least once reply must be posted before (midnight on Saturday).

15.2 Goals for the week:

1. Week 7 (Information) Review
2. Information Geometry
    - Example on Maximum Geometry

15.2.1 Challenge for the week:

Challenge: Use Maximum Geometry to prove the *Central Limit Theorem.*

15.3 Review of last week:

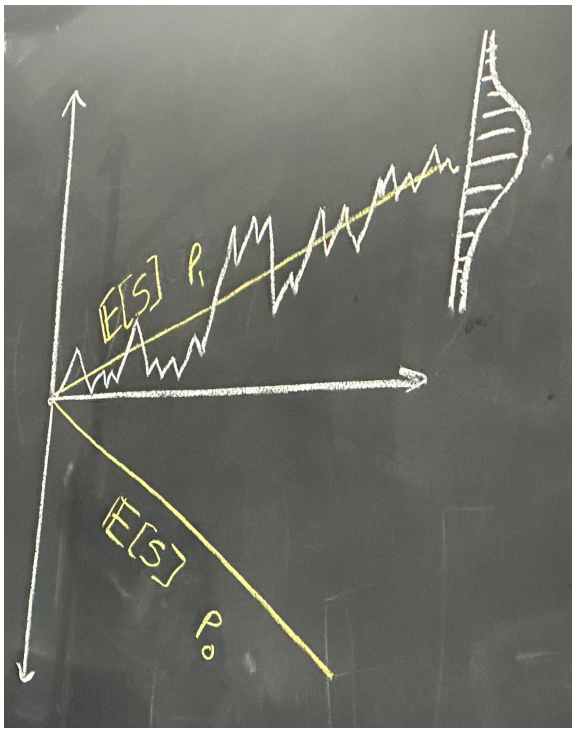
If we have $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$ where $p = p_0$ or p_1.

$$S = s(X_1, \ldots, X_n) = \log\left(\frac{\mathbb{P}[X_1, \ldots, X_n \mid p_1]}{\mathbb{P}[X_1, \ldots, X_n \mid p_0]}\right)$$

Note for quiz 4: The order for entries in a KL divergence do matter. You will be tested on this order. The one that goes first is the one you are drawing from:

$$\mathbb{E}_{X\sim p_1}[S] = nD(p_1||p_0) \tag{1}$$

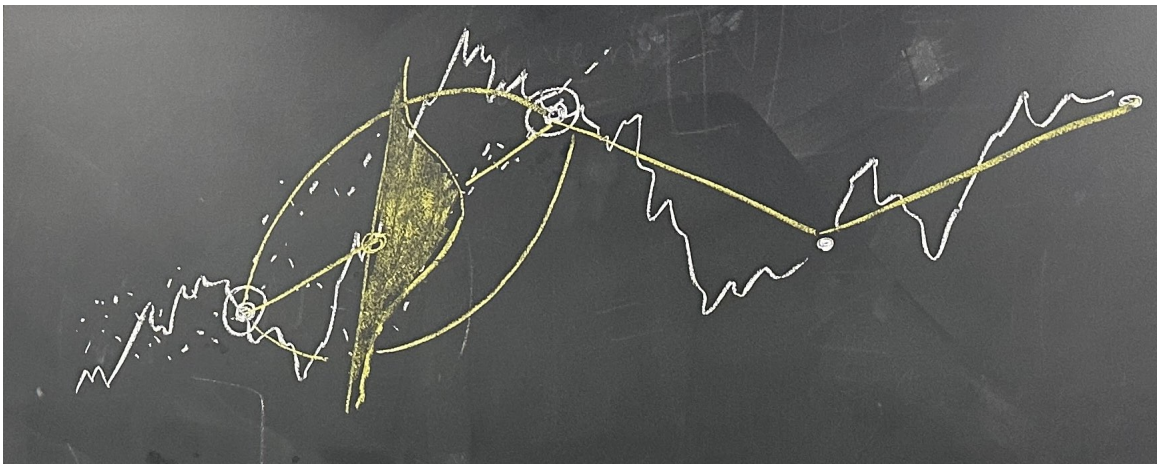
$$\mathbb{E}_{X\sim p_0}[S] = -nD(p_0||p_1) \tag{2}$$



Note that D controls the slope: increasing it (a change in D) tells you that it will be more clearly distinguishable.

Vertical distribution = background sampling distribution.

15.4 Brownian Bridge



15.5 Problem:

1. Target distribution p_0

2. Chosen divergence D
3. Variational family of distribution $\mathcal{P} = \{$ set of dist. $\}$
   (i) implicit: via constraints
   (i) explicit: via parameterized model $\{P_\Theta\}_{\text{all } \Theta}$

Goal: minimize $D(p||p_0), D(p_0||p)$ given: $p \in \mathcal{P}$.

The class of F-divergences is convex, so it can be formulated as a convex optimization problem in most cases.

15.5.1 Variational Calculus

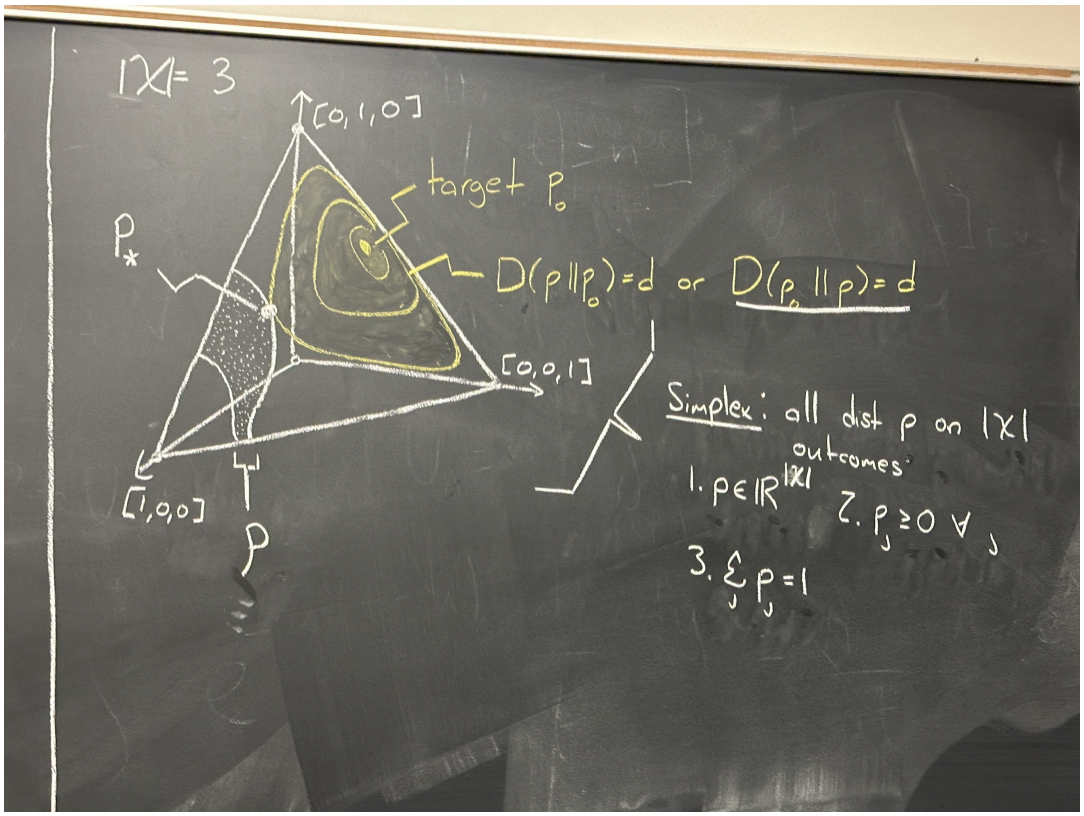
Note that you will need to use variational calculus instead of regular calculus in some choices of divergence (see the book for examples on this).

15.5.2 Fréchet derivatives

The derivative defined on normed spaces.

15.5.3 3D Probability Simplex

Within this set is our target – note that any point in the set is a distribution.



We look at the level sets that have distance d away from our target p_0.

15.6 Convergence in distribution on a Weak Topology

$\underbrace{D_{KL}(p_0||p)}_{\to 0} \geq \frac{1}{\ln 2} \underbrace{\mathrm{TV}(p, p_0)^2}_{0} \to p \Rightarrow p_0$ which is weak convergence, where TV is the total variation distance.

15.7 Maximum Entropy: A constrained optimization problem

$X \sim p$, where p is unknown, but we do "know" some moments.

We implicitly define $\mathcal{P}$ as follows:

15.7.1 Equality:

Equality $\mathcal{F}_i[p] = \mathbb{E}_{X\sim p}[f_i(x)] = \{\alpha_i\}_{i=1}^n$.

Examples:

$$\mathbb{E}[X], f(x) = x$$
$$\mathbb{E}[X^2], f(x) = x^2$$
$$\vdots$$

15.7.2 Inequality:

Inequality $\mathcal{G}_i[p] = \mathbb{E}_{X\sim p}[g_j(x)] \geq \{\alpha_j\}_{j=1}^m$.

15.7.3 Goal

We want to $\max H(p)$, $\quad p \in \mathcal{P} \implies p_* = \text{argmax}_{p\in\mathcal{P}}\{H(p)\}$.

15.8 Question:

$$\max_{p\in\mathcal{P}}\{H(p)\} \iff \min_{p\in\mathcal{P}}\{D(?||?)\}? \tag{3}$$

! Quiz: You want to relate the above to the KL divergence of a uniform distribution.

$$\begin{aligned}
H(p) &= -\mathbb{E}_{x\sim p}[\log(p(x))] = -\mathbb{E}_{x\sim p}[\log(\frac{p(x)}{1/|\mathcal{X}|}\cdot\frac{1}{|\mathcal{X}|})] \\
&= -\mathbb{E}_{x\sim p}[\log(\frac{p(x)}{1/|\mathcal{X}|})] + \mathbb{E}_{x\sim p}[\log(|\mathcal{X}|)] \\
&= -D(p||U[\mathcal{X}]) + H[U[\mathcal{X}]] \\
H(p) &= H(U(\mathcal{X})) - D(p||U[\mathcal{X}]) \\
\implies \max_{p\in\mathcal{P}}\{H(p)\} &\iff \underbrace{\min_{p\in\mathcal{P}}\{D(p||U[\mathcal{X}])\}}_{\text{Occam's Razor for } \max H}
\end{aligned}$$

15.9 Karush-Kuhn-Tucker (KKT) Conditions:

For Convex differentiable functions, how do we find sufficeient and necessary stationary conditions? How do we generalize Lagrange multipliers for when we have inequality constraints?

$$\text{Given } \mathcal{P} = \left\{ p \text{ s.t. } \underbrace{\{\mathcal{F}_i[p] = \alpha_i\}_{i=1}^n}_{\text{equality}}, \underbrace{\{\mathcal{G}_j[p] = \alpha_j\}_{j=1}^n}_{\text{inequality}} \right\}$$

For an objective: $H(p)$

1. $-\nabla_p H(p) + \sum_{i=1}^{n} \lambda_i \nabla_p \mathcal{F}(p) + \sum_{j=1}^{n} \mu_j \nabla_p \mathcal{G}(p)\Big|_{p=p^*} = 0$
2. $p^* \in \mathcal{P}$
3. $\mu_j \geq 0 \quad \forall j$
4. if $G_j[p] \gneq \beta_j$ then $\mu_j = 0$.

moment: $\mathcal{F}[p] = \mathbb{E}_{x\sim p}[f(x)] = \sum_{\text{all } x} p(x) f(x)$

We are restricted to a polytope, that is a subset of a face on the simplex.

The constraints listed above are called:

0th order optimality condition: $\forall x, p(x) \geq 0 \mapsto p(x') = \begin{cases} 0 & \text{if } x' \neq x \\ 1 & \text{if } x' = x \end{cases}$.

1. Stationarity (first order optimality condition): $\sum_x p(x) = 1 \mapsto \partial_{p(x)} \sum_{x'} p(x') = 1$.
2. Feasibility (Primal), it has to live inside the set of possible distributions.
3. Feasibility (Dual): Non-negativity, your gradient should be pointed out (not into) your set.
4. Complementary Slackness (Active/Inactive): $\lambda_i^* h_i(x^*) = 0, \forall i \quad \iff \sum()$.

The gradient of our objective function must be parallel to the gradient of our constraints.

15.10 Taking Derivatives:

$$\begin{aligned} H(p) &\mapsto -\partial_{p(x)} \mathbb{E}_{x\sim p}[\log(p(x))] \\ &= -\partial_{p(x)} \sum_{x'} [p(x') \log(p(x'))] \\ &= -\partial_{p(x)} [p(x) \log(p(x))] \\ &= -\log(p(x)) - 1 \\ &= -(\log(p(x)) + 1) \end{aligned}$$

If the normal distribution maximizes entropy for a known mean and variance, since it is capped above by a gaussian's .

Convergence in KL is convergence in distribution.

16 Thursday, March 7th

16.1 Method of Types

- $X \in \mathcal{X}, |\mathcal{X}| < \infty$
- $X^{(n)} = \{X_i\}_{i=1}^n, X \overset{\text{i.i.d.}}{\sim} p_0$

16.2 Empirical Distribution

$X^n = x^n$ is

- Draw $J \sim \text{Uniform}[1, n]$
- Y = X_j

$P(Y = y) = \frac{\#\text{ of s.t. } x_j = y}{n}$

$$\mathcal{P}^{(n)} = \{\text{frequency of the outcome } y \text{ among } \{x_i\}_{i=1}^n\} \tag{4}$$

16.3 Type

Definition: A type P is a probability distribution on $|\mathcal{X}|$ outcomes w/ rational prob., denominator is n.

16.3.1 Type Class

The type-class $T(P)$,

$$T(P) = \{x^{(n)} \text{ s.t. empirical dist } P_{X^{(n}}(\cdot) = P\} \tag{5}$$

1. The size of $T(P)$ is combinatorial, # distinct rearrangements of $x^{(n)}$ of $T(P_{x^{(n)}})$
2. all draws $x^{(n)} \in T(P)$ are equiprobable

Looking at draws (type class) of length 6, when the event can have 3 outcomes, we view them as points in the simplex (space of possible distributions) as opposed to bar charts.

16.4 Theorem 11.1.1

As stated in T&C, The number of types with denominator n is bounded by

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

16.5 Theorem 11.1.2

As stated in T&C, If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. according to $Q(x)$, the probability of x^n depends only on its type and is given by

$$Q^n(x^n) = 2^{-n(H(P_x n)+D(P_x n||Q))}.$$

16.5.1 Multiplicity vs Fitting to Data Generating Distribution

Multiplicity: $H_d(P_{x^{(n)}})$.

Fitting to Data Generating Distribution p_0: $D_d(P_{x^{(n)}}||p_0)$

16.6 Theorem 11.1.3: Size of a type class T(P)

For any type $P \in \mathcal{P}_n$,

$$\underbrace{\frac{1}{(n+1)^{|\mathcal{X}|}}}_{\text{Polynomial}} d^{nH(P)} \leq |T(P)| \leq d^{nH(P)}.$$

This goes beyond excess surprise and gives us something to bound any probability distribution by.

16.7 Theorem 11.1.4: Prob. of a type class T(P)

For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P\|Q)}$ to first order in the exponent. More precisely,

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

16.8 Key Equation

$$D(P\|Q) \overset{n\to\infty}{\simeq} = -\frac{1}{n}\log_d(P(X^{(n)} \in T(P)).$$

where $\simeq$ means asymptotically equal. This is used in several proofs, including that of CLT.

17 Tuesday, March 12th

17.1 Logistics

- Project Part 3 due today – Part 4 will be out this week
  - Note you only have finitely many slip days.
- Discussion due Thursday
- Quiz 4 retake Thurs (9:30am, 1:30pm)
- Quiz 5 (on chapter 11) on Thursday
  - Note this plays a similar role as Chapter 5 on Entropy, but on Information: Interpreting info in more applied settings. Thus this will be mainly a quiz on statements/theorems.

17.2 Activity Groups:

We will examine the following 3 so we can determine proper use cases between them. The distinguishability (expectation of the test statistic) of two distributions is its KL Divergence.

$D(p\|q)$ is how distinguishable p is from q, where distinguishabililty is how much samples you need asymptotically, s.t. you can separate them (tell which came from which distribution) w.h.p..

$I[X;Y]$ is $D(p_{X,Y}\|p_X p_Y)$, how distinguishable the true joint is from the product (thus they are a measure of dependence). Note that this even holds for non-linear dependence.

Example:

$H(p)$ is the compression of a r.v..

17.2.1 LLN (Thm 11.2.1)

Theorem 11.2.1 Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim P(x)$. Then

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

and consequently, $D(P_{x^n}\|P) \to 0$ with probability 1.

Proof: The inequality (11.69) was proved in (11.68). Summing over n, we find that

$$\sum_{n=1}^{\infty} \Pr\{D(P_{x^n}\|P) > \epsilon\} < \infty$$

Thus, the expected number of occurrences of the event $\{D(P_{x^n}\|P) > \epsilon\}$ for all n is finite, which implies that the actual number of such occurrences is also finite with probability 1 (Borel-Cantelli lemma). Hence $D(P_{x^n}\|P) \to 0$ with probability 1. □

The Empirical Distribution converges i.p. to the true distribution as $n \to \infty$.

The prob. of observing an Empirical Distribution that is far away from the true distribution is vanishing (so we converge i.p.).

Thus w.h.p., $D(p\|q) < \varepsilon$.

Look for $\mathcal{E}(n)$ $n \overset{\to}{\to} \infty$ 0 s.t. $\mathbb{P}[D(P_{x^n}\|P_0) < \mathcal{E}(n)] \to 1$.

This leads to the fact that we can do estimation with data.

17.2.2 Sanov & Conditional Limit (Thm 11.4.1, 11.6.2)

Theorem 11.4.1 (Sanov's theorem) Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then

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

where

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

is the distribution in E that is closest to Q in relative entropy. If, in addition, the set E is the closure of its interior, then

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

Proof: We first prove the upper bound:

$$\begin{aligned}
Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P\|Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P\|Q)} \\
&= \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E \cap \mathcal{P}_n} D(P\|Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E} D(P\|Q)} \\
&= \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P^*\|Q)} \\
&\leq (n+1)^{|\mathcal{X}|_2} 2^{-nD(P^*\|Q)},
\end{aligned}$$

where the last inequality follows from Theorem 11.1.1. Note that P^* need not be a member of $\mathcal{P}_n$. We now come to the lower bound, for which we need a "nice" set E, so that for all large n, we can find a distribution in $E \cap \mathcal{P}_n$ that is close to P^*. If we now assume that E is the closure of its interior (thus, the interior must be nonempty), then since $\cup_n \mathcal{P}_n$ is dense in the set of all distributions, it follows that $E \cap \mathcal{P}_n$ is nonempty for all $n \geq n_0$ for some n_0. We can then find a sequence of distributions P_n such that $P_n \in E \cap \mathcal{P}_n$ and $D\left(P_n\|Q\right) \to D\left(P^*\|Q\right)$. For each $n \geq n_0$,

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

Consequently,

$$\begin{aligned}\liminf \frac{1}{n} \log Q^n(E) &\geq \liminf \left(-\frac{|\mathcal{X}| \log(n+1)}{n} - D\left(P_n || Q\right) \right) \\ &= -D\left(P^* || Q\right).\end{aligned}$$

Combining this with the upper bound establishes the theorem. □

This argument can be extended to continuous distributions using quantization.

Sanov: Limits & bounds on P[a (sort of) extreme event] that converges to 0 (as $n \to \infty$), exponentially with $d^{-n \cdot D(p_* \| p_0)}$ (in n w/ role $D(p_* \| p_0)$).

Conditional: Conditioned on $P_{x^n} \in E$.

$$P_{x^n} \overset{\text{i.p.}}{\to} p_*$$

17.2.3 Hypothesis Testing (Chernoff-Stein Thm 11.8.3)

Theorem 11.8.3 (Chernoff-Stein Lemma) Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q$. Consider the hypothesis test between two alternatives, $Q = P_1$ and $Q = P_2$, where $D(P_1 \| P_2) < \infty$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for hypothesis H_1. Let the probabilities of error be

$$\alpha_n = P_1^n(A_n^c), \quad \beta_n = P_2^n(A_n).$$

and for $0 < \epsilon < \frac{1}{2}$, define

$$\beta_n^\epsilon = \min_{\substack{A_n \subseteq \mathcal{X}^n \\ \alpha_n < \epsilon}} \beta_n.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1 \| P_2).$$

Inuitively, if we allow α to be fixed, then $P_\lambda^* = P_0$ (exponent does not decay) and hence $\beta \approx 2^{-nD(P_0\|P_1)}$, i.e. we can achieve a faster error exponent on one type of error probability if we allow the other type of error probability to be fixed or decay arbitrarily slowly. For a rigorous proof, see below:

Proof: We prove this theorem in two parts. In the first part we exhibit a sequence of sets A_n for which the probability of error β_n goes exponentially to zero as $D(P_1 \| P_2)$. In the second part we show that no other sequence of sets can have a lower exponent in the probability of error.

For the first part, we choose as the sets $A_n = A_\epsilon^{(n)}(P_1 \| P_2)$. As proved in Theorem 11.8.2, this sequence of sets has $P_1(A_n^c) < \epsilon$ for n large enough. Also,

$$\lim_{n\to\infty} \frac{1}{n} \log P_2(A_n) \leq -(D(P_1 \| P_2) - \epsilon)$$

from property 3 of Theorem 11.8.2. Thus, the relative entropy typical set satisfies the bounds of the lemma.

To show that no other sequence of sets can to better, consider any sequence of sets B_n with $P_1(B_n) > 1 - \epsilon$. By Lemma 11.8.1, we have $P_2(B_n) > (1 - 2\epsilon)2^{-n(D(P_1\|P_2)+\epsilon)}$, and therefore

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log P_2(B_n) &> -(D(P_1 \| P_2) + \epsilon) + \lim_{n\to\infty} \frac{1}{n} \log(1 - 2\epsilon) \\ &= -(D(P_1 \| P_2) + \epsilon). \end{aligned}$$

Thus, no other sequence of sets has a probability of error exponent better than $D(P_1 \| P_2)$. Thus, the set sequence $A_n = A_\epsilon^{(n)}(P_1 \| P_2)$ is asymptotically optimal in terms of the exponent in the probability. □

Less formally, that is:

We are drawing $X_{i i=1}^n \overset{\text{i.i.d.}}{\sim} p$, where $p = p_1$ or $p = p_2$.

We have two probabilities of errors that we need to control for.

α_n = FNR = False Rejection. We control this by forcing it to be less than $\varepsilon \in [0, \frac{1}{2}]$ with the exact value acting as a bound on an error rate guarantee.

β_n = FPR = False Acception.

β_n^ε = Best you can do for one error, as you control for the other error ($\alpha_n < \varepsilon$).

then: the min(FP) w/ which we can distinguish p_1 and p_2 is exponential in n: proportional to $d^{-nD(p_2\|p_1)}$ for the best possible test.

As we make n bigger, we see that it grows roughly exponential in the number of samples: $e^{nD(P(P_1\|P_2))}$.

If this decays to 0 quickly, then you only need a few samples.

If this decays to 0 slowly, then you will need a lot of samples.

This is measured via the Divergence of the distributions.

This rate of convergence, the further apart distributions are, the easier it should be to tell they are different.

This is done as a fn. of how much data I collect.

18 Thursday, March 14th: Practical Problems in Information Geometry

18.1 Logistics

1. Peer Review Assigned (due next Thurs.)
2. Discussion Due Today
3. Quiz 5 Today

18.2 Common Questions

1. Choice of variational class
2. Choice/Interpretation of Loss
3. Estimating (from data)
   - The loss ($D, I, H \implies$ need density)
   - ∇_p of the loss

18.3 Canonical Questions

1. Fix $\mathcal{X}$
2. Fix target P_0
3. Choose variational family $\mathcal{P}$
4. Choose loss: $D(\cdot\|p_0)$ or $D(p_0\|\cdot)$, which we note is cvx.

Find

$$p_* = \text{argmin}_{p\in\mathcal{P}} \{\text{loss}(p, p_0)\} \tag{6}$$

You can still run into issues if the class you are restricting to (on the 3D simplex) is not a convex class.

18.4 Variational Bayes

Usually in a Bayesian setting, you want some function (i.e. mode, or even a credible interval),

$$p_0 = \text{posterior distribution}$$

If we don't have a conjugate prior, we often ignore the normalization factor (the denominator) when we take the product of the likelihood and the prior. Note: we could find it if we *really* wanted to, but integration in high dimensions is hard.

So we only know p_0 up to a normalization factor.

This means we know $\log(p_0) + C$ up to an unknown additive factor, C.

We can sample from this, but as everyone in the class already knows, sampling is very difficult. So in Variational Bayes, you often find a tractable family:

$$\mathcal{P} = \{\text{set of "tractable" distributions, sufficiently expressive}\} \implies X^{(n)} = \{X_1\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p \quad \text{if } p \in \mathcal{P} \tag{7}$$

Our loss is $D(p\|p_0)$, restricting $p \in \mathcal{P}$. We minimize this loss over all $p \in \mathcal{P}$.

18.4.1 Mean-Field Approximation

This is implicitly given from Variational Bayes as it simply a form of Block Independence (between entries of $X \in \mathcal{X}$ where $X = [X_1, X_2, \ldots, X_n]$ is a vector and $\mathcal{X}$ the space of possible outcomes).

In here, we get explicit parametric forms for the solution $p_* \propto \exp(\sum_{i=1}^n \lambda_i g_i(x))$ which we optimize with the Expectation-Maximization algorithm.

18.4.2 Gaussian Mixture Model

$$\mathcal{P} = \{P_\theta\}_\Theta \tag{8}$$

18.5 Generative Models

1. Know $\{X_1\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p_0$, for unknown p_0, $P_{x^{(n)}} \approx p_0$
2. $\mathcal{P} = \{p_\theta\}_\Theta$, which we can optimize over with a NN
3. Loss: $D(P_{x^{(n)}}\|p_\theta)$, $p \in \mathcal{P} = \{p_\theta\}_\Theta$
   - average over $\{X_i\}_{i=1}^n$

   This is equivalent to MLE due to the method of types (we optimize over Θ).

18.5.1 Proof of Generative Models having strong duality with MLE

$\mathcal{L}(\theta; X^{(n)}) = d^{-n(H(P_{x^{(n)}})+D(P_{x^{(n)}}\|P_\theta))}$ is equivalent to: $d^{-n(D(P_{x^{(n)}}\|P_\theta))}$

which is equivalent to: $\min n(D(P_{x^{(n)}}\|P_\theta))$

18.5.2 Large Language Models

In LLMs, we formulate this as minimizing "Perplexity" or "Cross-Entropy" which is equivalent to:

$$\begin{aligned}\min(D(P_{x^{(n)}}\|P_\theta)) &\equiv \min \mathbb{E}_{Y\sim P_x^{(n)}}\left[\log\left(\frac{P_x^{(n)}(Y)}{p_\theta(Y)}\right)\right] \\ &\sim -\mathbb{E}_{Y\sim P_x^{(n)}}\left[\log\left(p_\theta(Y)\right)\right]\end{aligned}$$

where $P_x^{(n)}(Y) = \begin{cases} 0 & \text{if } X_i \neq y \\ \frac{1}{n} & \text{o/w} \end{cases}$

If you get your distribution wrong, you will be more surprised than if you were correct about the distribution it came from: $H[p,q] \geq H[p]$. We can also see this by looking at the difference.

18.5.3 Proof via revisiting previous lectures (incorporate KL Divergence)

$$\begin{aligned} H(p,q) - H(p) &= \mathbb{E}_{X\sim p}[-\log(q(X)) + \log(p(X))] \\ &= \mathbb{E}_{X\sim p}[-\log(\frac{p(X)}{q(X)})] \\ &= D(p\|q) \\ &\geq 0 \qquad \text{[Since divergences are non-negative]} \end{aligned}$$

Thus we get that $H(p,q) - H(p) \geq 0 \implies H(p,q) \geq H(p)$

18.5.4 Relating back to Perplexity

We want to $\min D(p\|q)$ w.r.t. $q : D(p\|q) = \underbrace{H(p,q)}_{\min H(p,q) \implies} - H(q)$

$$\begin{aligned} \implies \text{Perplexity of } p_\theta, X^{(n)} &= d^{-H(p_x^{(n)}, p_\theta)} \\ &= d^{-\frac{1}{n}\sum_{i=1}^{n} \log(p_\theta(x_j))} \end{aligned}$$

Most Generative models give non-zero probability to any possible next word. However this leads to extremely limited support. Since we are restricted to samples makes it so that we don't run into issues with the non-symmetric aspect of the KL divergence $D(p_\theta \| P_x^{(n)}) = \infty$.

18.6 Privacy/Fairness/Data Augmentation

True joint dist. p_0,

$$(\underbrace{X}_{\text{data observe}}, \underbrace{Y}_{infer}, \underbrace{Z}_{protect}) \sim p_0$$

Look for data-processing: $T(X) \to X'$ given $T \to p' \sim (X', Y, Z)$.

Goal: choose T to:

1. $\max_T \{I[X';Y]\}$
2. $\min_T \{I[X';Z]\}$

19 Tuesday, March 19th

19.1 Logistics

1. Peer Review – due Thursday.
2. Chapter 4 Reading posted
3. Quiz 5 retake will be on Wednesday 5:00pm

19.2 Goals

- Bayes Optimal Experimental Design:

Information is expected Δ in dist. of unknown after observation.

19.3 Problem

- $X \in \mathbb{R}^d$ is some unknown
  - a priori: $X \sim p_0$
- have a set of questions $Q = \{q_j\}_{j=1}^m$ where, when we ask $q_j \to$ observe Y
  (Y is a measurement, it has noise)
- **each q is expensive**

- Aim: design a protocol for choosing an optimal sequence of questions... $\{q_{j(n)}\}_{n=1}^{\cdots}$
  - Greedy: choose $q_{j(n)}$ given the initial distribution for x (which is p_0) and everything we have learned so far: $Y^{(n-1)} = [Y_1, Y_2, \ldots, Y_{n-1}]$

We are often trying to find answers to continuous problems (Huffman was greedy in reverse – it worked by saving your question for the last 2 things – you build your tree backwards). Thus since there are an uncountable infinity amount of outcomes, we can't start backwards – we cannot use Huffman.

Since the Fano question is at most 2 questions (2 bits) worse than Huffman, we note that it is decent.

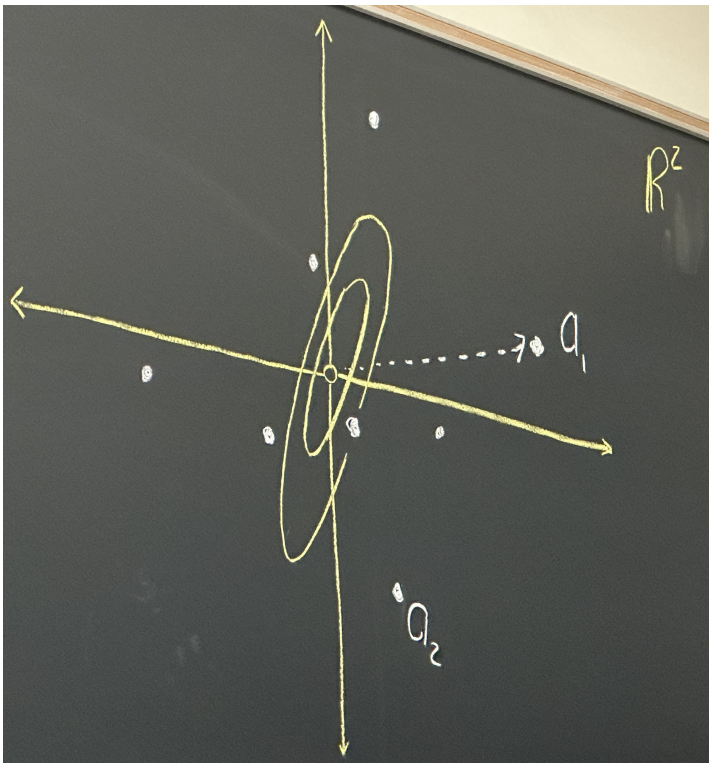


Figure 2: Ellipsoids

19.3.1 Criteria to optimize j(n)

1. We want to choose j at step n to minimize $\mathbb{E}_{Y_n=y_n}(H[X \mid y_n, Y^{(n-1)}]) = H[X \mid Y^{(n)}]$
2. We want to choose $j(n)$ to maximize $I[X; Y^n \mid Y^{(n-1)}]$.
3. We want to choose $j(n)$ to change your distribution the most, to maximize how much we expect the distribution to change (Information gain) after you ask the question

$$\mathbb{E}_{Y_n=y_n}[D(\underbrace{p_{X|y^n,Y^{(n-1)}}}_{p_{n|y_n}} \| \underbrace{p_{X|Y^{(n-1)}}}_{p_{n-1}})]$$

   where $p_{n|y_n}$ is the posterior and p_{n-1} is the prior.

We note that all 3 of the above criteria are equivalent. *Proof:*

1. Choose $j(n)$ s.t. it is the $\text{argmin}_k\{\mathbb{E}_{Y_n(k)=y_n(k)}(H[X \mid y_n(k), Y^{(n-1)}])\}$ over all kth questions being asked. Note that this is the same as $\text{argmin}_k\{H[X \mid Y_n(k), Y^{(n-1)}]\}$
2. $\text{argmin}_k\{H[X \mid Y^{(n-1)}] - H[X \mid Y_n(k), Y^{(n-1)}]\} = \text{argmin}_k\{I[X; Y_n(k) \mid Y^{(n-1)}]\}$ Interpretation: You are greedily trying to maximize the amount of information gained.
3. $$I[X; Y_n(k) \mid Y^{(n-1)}] = D(p_{X,Y_n(k)|Y^{(n-1)}} \| \underbrace{p_{X|Y^{(n-1)}}}_{p_{n-1}} p_{Y_n(k)|\cancel{Y^{(n-1)}}})$$

   We started out by writing out mutual information as a KL divergence between the joint and the product of the marginals. But we note that the second marginal distribution in the

product (2nd param of the KL Div.) is independent and thus can be removed. Thus we get:

$$
\begin{aligned}
&= \mathbb{E}_{X,Y_n(k)|Y^{(n-1)}}\left[\log\left(\frac{p_{X,Y_n(k)|Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right] \\
&= \mathbb{E}_{Y_n(k)|\cancel{Y^{(n-1)}}}\left[\mathbb{E}_{X|y_n,Y^{(n-1)}}\left[\log\left(\frac{p_{X,Y_n(k)|Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right]\right] \\
&= \mathbb{E}_{Y_n(k)}\left[\mathbb{E}_{X|y_n,Y^{(n-1)}}\left[\log\left(\frac{p_{Y_n(k)|Y^{(n-1)}}p_{X|Y_n(k),Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right]\right] \\
&= \mathbb{E}_{Y_n(k)}\left[\mathbb{E}_{X|y_n,Y^{(n-1)}}\left[\log\left(\frac{p_{Y_n(k)}p_{X|Y_n(k),Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right]\right] \\
&= \mathbb{E}_{Y_n(k)}\left[\mathbb{E}_{X|y_n,Y^{(n-1)}}\left[\log\left(\frac{p_{Y_n(k)}p_{X|Y_n(k),Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right]\right] \\
&= \mathbb{E}_{Y_n(k)}\left[\mathbb{E}_{X|y_n,Y^{(n-1)}}\left[\log\left(\frac{p_{Y_n(k)}p_{X|Y_n(k),Y^{(n-1)}}}{p_{X|Y^{(n-1)}}p_{Y_n(k)}}\right)\right]\right]
\end{aligned}
$$

Thus since $I[X; Y_n(k) \mid Y^{(n-1)}] = \mathbb{E}_{Y_n(k)=y_n(k)}[D(\underbrace{p_{n,y_n}}_{p_{X|y_n,Y^{(n-1)}}} \quad \| \quad \underbrace{p_{n-1}}_{\text{posterior given } Y^{(n-1)}})]$.

Fun discussion problem: instead of an analytical solution as done below, utilize a decision tree to get an approximate solution.

Model:

$$X \sim \mathcal{N}(\mu, \Sigma_0)$$

each question $j : q_j : a_j^\top X + \varepsilon = Y(j)$, with $\varepsilon \sim \mathcal{N}(0, \sigma_j^2)$

for a fixed set of m questions: $\{q_j\}_{j=1}^m$.

$$q_j : Y(j) = a_j^\top X + \varepsilon \tag{9}$$

This gives you a confidence interval which with high probability, you expect the true X to lie when projected within the space – you are constraining the projection of X (with noise) onto a specific direction which changes the distribution.

Since every variable in this problem is gaussian and linear, we understand how the posterior updates after each observation.

Also, a Gaussian can be uniquely identified with 2 pieces of information: μ, σ. For a MVG, the entropy is solely dependent on the covariance matrix, so we simply need to observe how that updates.

Suppose: $p_{n-1}(X) = p(X \mid y^{(n-1)}) \sim \mathcal{N}(\mu_{n-1}, \Sigma_{n-1})$.
Observe: $Y_n(j) = y_n(j) = a_j^\top X + \varepsilon$.

Posterior:

$$\begin{aligned}
p_n(X) &= p(X \mid y_n(j), y^{(n-1)}) \\
&\propto \underbrace{p_{Y_n(j)}(y_n \mid x, \cancel{Y^{(n-1)}})}_{\text{given } x\text{: } Y_n(j)|x \sim \mathcal{N}(a_j^\top x, \sigma_j^2)} \; p(X = x \mid y^{(n-1)}) \\
&\propto \underbrace{\exp(-\frac{1}{2}\left(\frac{(y - a_j^\top x)^2}{\sigma_j^2} + (x - \mu_{n-1})^\top \Sigma_{n-1}^{-1}(x - \mu_{n-1})\right))}_{\dots}
\end{aligned}$$

where we took out the quadratic terms in x.

19.3.2 Rank 1 updates to the Precision Matrix

$$x^\top \underbrace{(\frac{1}{\sigma_j^2} a_j a_j^\top + \Sigma_{n-1}^{-1})}_{\Sigma_0^{-1}} x$$

where $\Sigma_0^{-1} \implies \Sigma_n = (\Sigma_{n-1}^{-1} + \frac{1}{\sigma_j^2} a_j a_j^\top)^{-1} \perp\!\!\!\perp Y^{(n)}$

and where $+\frac{1}{\sigma_j^2} a_j a_j^\top$ is a rank 1 update to the precision, $P = \Sigma^{-1}$.

In statistics, the precision matrix or concentration matrix is the matrix inverse of the covariance matrix or dispersion matrix, $P = \Sigma^{-1}$. For univariate distributions, the precision matrix degenerates into a scalar precision, defined as the reciprocal of the variance, $p = \frac{1}{\sigma^2} \cdot$.

19.3.3 Relating back to Entropy

$$\begin{aligned}
H[X \mid Y^{(n)}] &= \frac{d}{2} \ln(2\pi e) + \frac{1}{2} \ln(\det(\Sigma_n)) \\
&\mapsto I[X; y_n(j) \mid Y^{(n-1)}] \\
&= H[X \mid Y^{(n-1)}] - H[X \mid Y^{(n)}] \\
&= \frac{d}{2} \ln(2\pi e) + \frac{1}{2} \ln(\det(\Sigma_{n-1})) - \frac{d}{2} \ln(2\pi e) - \frac{1}{2} \ln(\det(\Sigma_{n|j(n)})) \\
&= \frac{1}{2} \ln(\frac{\det(\Sigma_{n-1})}{\det(\Sigma_{n|j(n)})})
\end{aligned}$$

Note that:

$$\det\left(\Sigma_{n-1}^{1/2}\Sigma_{n|j(n)}^{-1}\Sigma_{n-1}^{1/2}\right)$$
$$\det\left(\Sigma_{n-1}^{1/2}\left(\Sigma_{n-1}^{-1}+\frac{1}{\sigma_j^2}a_j a_j^\top\right)\Sigma_n^{1/2}\right)$$
$$\det\left(I+\frac{1}{\sigma_j}\left(\Sigma_n^{1/2}a_j\right)\left(\Sigma_n^{\frac{1}{2}}a_j^\top\right)\right)$$

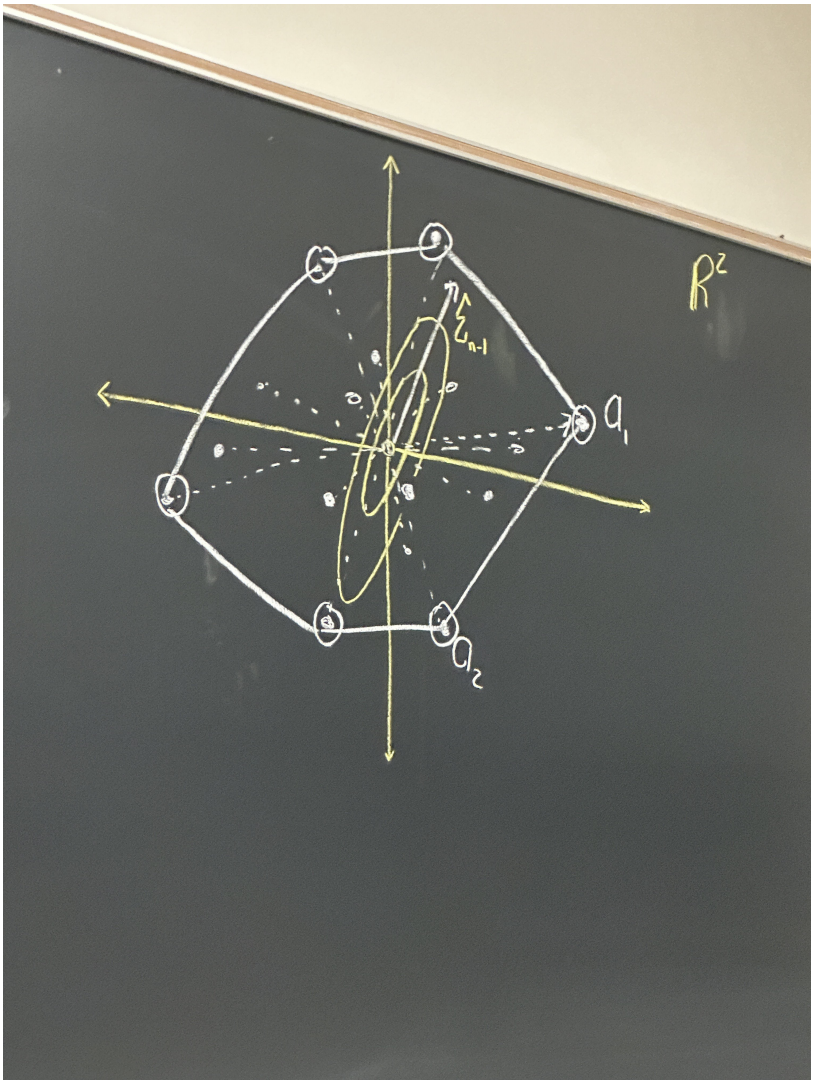


Figure 3: Constrained Ellipsoid

19.4 Quadratic Programming is a Geometry Problem

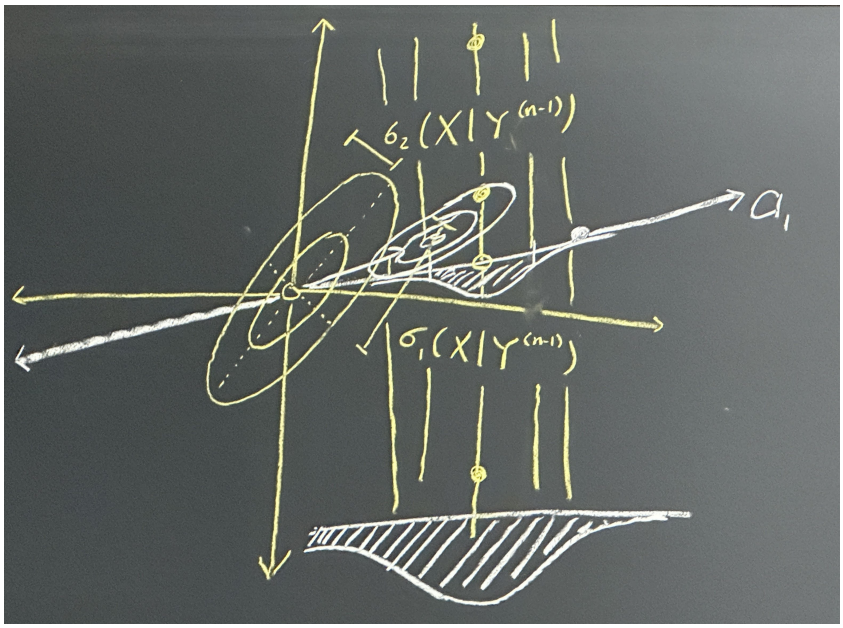


Figure 4: Quadratic Program

$$\lambda_i(I + A) = 1 + \lambda_i(A)$$

$$\left(uu^\top\right) = \begin{bmatrix} \\ u \\ \\ \end{bmatrix} \begin{bmatrix} u^\top \end{bmatrix}$$

- Acquisition: pick $j(n) = \text{argmax}_k\{\frac{1}{\sigma_k}^2 a_k^\top \Sigma_{n-1} a_k\}$
- Quadratic programming on a discrete set

$$\left(uu^\top\right) u = u\left(u^\top u\right) = \underbrace{\|u\|^2 u}_{\lambda} \tag{10}$$

20 Thursday, March 21st: Stochastic Processes

20.1 Logistics

1. Peer Review – due Today.
2. Optional Reading posted – T&C 4, GP notes.
3. Enjoy Spring Break!

20.2 Goals

- (·) Processes
- (·) Entropy Rates

20.3 Stochastic Process

A **stochastic process** $\{X_j\}_{j=1}^{\cdots}$ or $\{X(t)\}_t$ is indexed by "time", and is discrete or continuous.

20.3.1 Joint distribution over a countable random vector

There will be a joint distribution for any subset ($\{X_j\}_{j\in\mathcal{J}}$ or $\{X_j\}_{j\in\mathcal{J}}$) of variables, we have:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{11}$$

There is no notion of 0 time – it's arbitrary.

20.3.2 Stationary of a Process

A process is **stationary** if $p(\{X_j\}_{j\in\mathcal{J}} = \{x_j\}) = p(\{X_{j+s}\}_{j\in\mathcal{J}} = \{x_j\})$, if any set of samples is unchanged under time translation.

20.3.3 Stationary Distribution

If a process, $\{X_j\}_{j=1}^{\cdots}$, is stationary then the corresponding Stationary Distribution is:

$$p_{\text{stat}}(x) = P(X_j = x) \text{ for any } j \tag{12}$$

20.3.4 Ergodicity of Stochastic Processes

If a process, $\{X_j\}_{j=1}^{\cdots}$, is stationary then it is also Ergodic:

$$\mathbb{E}_{X\sim p_{\text{stat}}}[f(X)] = \bar{f}_{\text{stat}} \tag{13}$$

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) = \bar{f}_{\text{time}}(n) \overset{n\to\infty}{\to} \bar{f}_{\text{time}} \tag{14}$$

We expect the long-term avg to be concentrated and to converge.

The fact that (13) and (14) are the same, $\bar{f}_{\text{stat}} = \bar{f}_{\text{time}}$, is what makes us ergodic.

20.4 Markov Process

A random process whose future probabilities are determined by its most recent values. A stochastic process $X(t)$ is called Markov
if $X(t+s), \quad s>0$ is conditionally $\perp\!\!\!\perp$ of $X(t-h), \quad h>0$ given $X(t)$.

The future is conditionally independent of the past given the present.

$$\Pr\left(X_{n+1}=x_{n+1} \mid \underbrace{X^{(n)}}_{X_n, X_{n-1}, \ldots X_1} = x^{(n)}\right) = \underbrace{P\left(X_{n+1}=x_{n+1} \mid X_n = x_n\right)}_{\text{Transition prob.: } P_{i,j}^{(n)} = P(X_{n+1}=i \mid X_n = j)} \tag{15}$$

20.4.1 Time Autonomous

$$\Pr\left(X_{n+i}=i \mid X_n = j\right) = \Pr\left(X_2 = i \mid X_1 = j\right) \quad \forall i,j \text{ and } n. \tag{16}$$

20.5 Transition Matrix

The matrix P has entries $P_{i,j}$ which give you the probability of going from j at the present to i in the future. That is, $P_{i,j} = \Pr(X_{n+1} = i \mid X_n = j)$, $\quad P = \begin{bmatrix} | & | & & | \\ p_1 & p_2 & \cdots & p_{|\mathcal{X}|} \\ | & | & & | \end{bmatrix}$ with the jth column

being the conditional distribution of the future given the present.

! Note that we use the following convention: this matrix is column-stochastic.

$$\Pr(X^{(n)} = x^{(n)}) = \underbrace{\left(\prod_{j=1}^{n} \mathbb{P}_{X_{j+1}X_j}\right)}_{\Pr(X^{(n)}=x^{(n)}|X_1=x_1)} p(X_1 = x_1) \tag{17}$$

where $p(X_1 = x_1)$ is the initial distribution for the probability of where I start.

20.6 Irreducibility

A discrete-time, discrete-state MC is irreducibile if $\forall x_1 = i, x_n = j \quad \exists$ a path $x^{(n)} = \{i, \ldots, j\}$ s.t. $P(X^{(n)} - x^{(n)} \mid X_1 = i) > 0$.

20.7 Aperiodic

A discrete-time, discrete-state MC is aperiodic if it's period is 1.

20.7.1 Period

A Period of the process is its GCD of the length of all cycles:

$$x^{(n)} = \{x_1, x_2, \ldots, x_1\}, \quad P(X_n = x_n \mid X_1 = x_1) > 0$$

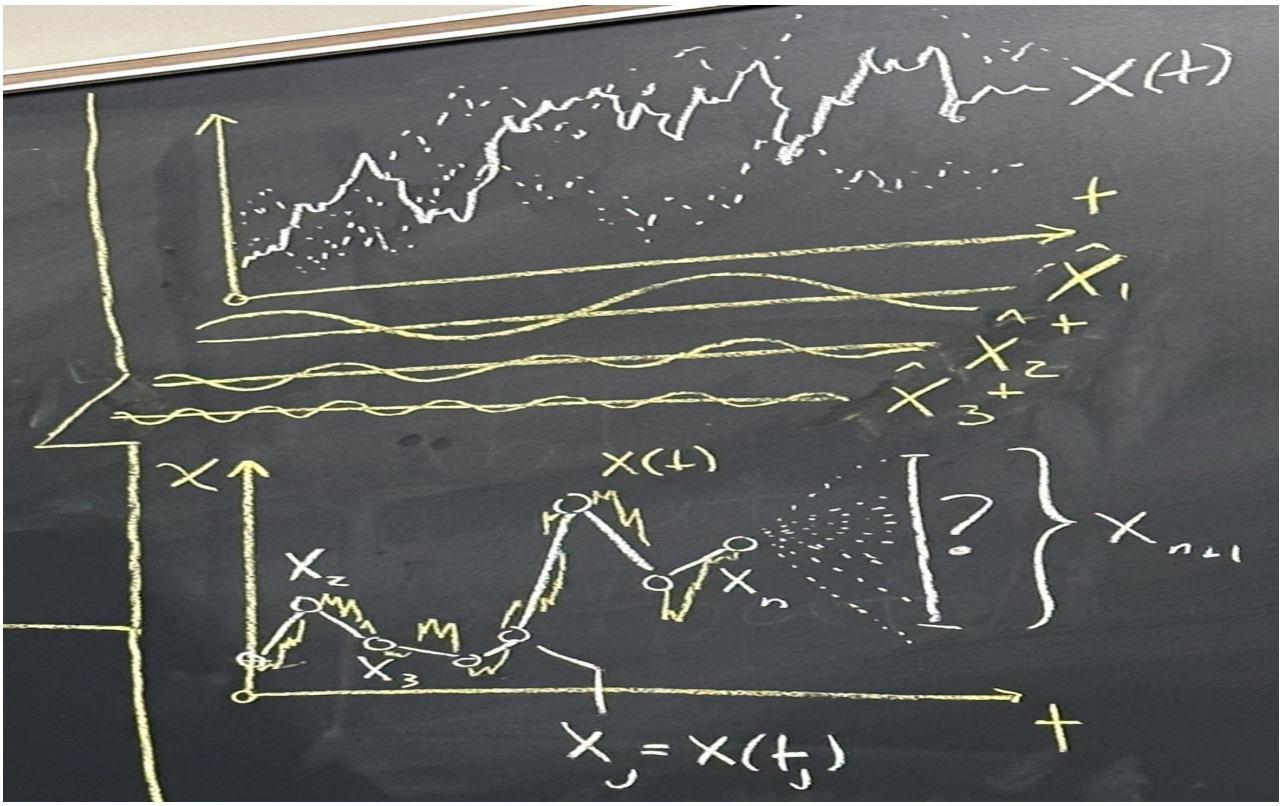


Figure 5: Various periods of Stochastic Processes

20.8 Perron-Frobenius Theorem

If P is column-wise Stochastic (transition for a $d-t, d-s$ MC) that is irreducible & aperiodic then:

it converges to a unique stationary distribution from any initial distribution, is stationary if initialized with P_{stat}, is ergodic.

1. $\exists$ a simple (A simple eigenvalue is an eigenvalue with multiplicity one – the eigenvalue is unique) largest (in magnitude) eigenvalue $\lambda_{\max}(P) \in \mathbb{R}^{+}$.
2. It's corresponding eigenvector $v_{\max}(P) \quad v_{\max}(x) \geq 0$ $(v \in \mathbb{R}^{+|\mathcal{X}|})$

Tldr;
A real square matrix with positive entries has a unique eigenvalue of largest magnitude and that eigenvalue is real ($\lambda \in \mathbb{R}$).

20.9 Gershgorin Disk Theorem

If $A \in \mathbb{C}^{n \times n}$ (or for simplicity, we will just provide a proof for $A \in \mathbb{R}^{n \times n}$), then $\lambda_j(A) \in \cup_{j=1}^{n} D(a_{jj}, \sum_{i \neq j} |a_i|)$

$\sum_{i \neq j} |p_{ij}| = \sum_{i \neq j} p_{ij} = 1 - p_{jj}$

This gives us that $|\lambda_j(P)| \leq 1$.

$\therefore \quad \lambda_{\max}(P) = 1$

$|\lambda_j(P)| = 1 \quad$ (if $j \geq 2$).

Algorithm 1 Algorithm for $p_0 : p_n \overset{n\to\infty}{\to} \mathbb{R}$

Initialize: $\Pr(X_0 = x) = p_0$.

Find: $p_n^{(x_n)} = \Pr(X_n = x_n) = \sum_{x_{n-1}} \underbrace{\Pr(X_n = x_n \mid X_{n-1} = x_{n-1})}_{P_{x_n, x_{n-1}}} \underbrace{\Pr(X_{n-1} = x_{n-1})}_{P_{n-1}^{(x_{n-1})}}$

Ensure: Matrix columns add to 1.

while stopping criteria not met **do**

$p_n \leftarrow P p_{n-1}$

$p_n \leftarrow P^n p_0 = V \Lambda^n V V^\top p_0 = V_{\max}(V V_{\max}^\top p_0) + \mathcal{O}(\lambda_2^n) = V_{\max}(\underbrace{\mathbb{1}^\top p_0}_{1}) + \mathcal{O}(\lambda_2^n) = V_{\max} + \mathcal{O}(\lambda_2^n)$

end while

return p_n

20.10 Entropy Rate

Define $H'[\mathcal{X}] = \lim_{n\to\infty} H[X_n \mid X^{(n-1)}]$.

1. If the process is stationary then the sequence $H[X_n \mid X^{(n-1)}]$ is non-increasing and converges.

Proof. Undoing a conditioning should increase entropy:

$$\begin{aligned}
H[X_n \mid X^{(n-1)}] &= H[X_n \mid X_{n-1}, X_{n-2}, \ldots, X_2, X_1] \\
&\leq H[X_n \mid X_{n-1}, X_{n-2}, \ldots, X_2] \\
&= H[X_{n-1} \mid X_{n-2}, X_{n-3}, \ldots, X_1] \qquad \text{[Invariant under time shifts]} \\
&= H[X_{n-1} \mid X^{(n-2)}] \\
\underbrace{H[X_n \mid X^{(n-1)}]}_{>0} &\leq H[X_{n-1} \mid X^{(n-2)}] \qquad \text{[non-incr seq + bounded below } \implies \text{ conv.]}
\end{aligned}$$

□

2. The limit of the new entropy (the cndtl entropy in the future given the past) is $H[\mathcal{X}] = H'[\mathcal{X}]$.

Proof.

$$H[\mathcal{X}] = \lim_{n\to\infty} \frac{1}{n} H[X^{(n)}] = \lim_{n\to\infty} (\frac{1}{n} \sum_{j=1}^{n} H[X_j \mid X^{(j-1)}]) \to H'[X]$$

□

This can also be understood (the difference between the observed value of a variable at time t and the optimal forecast of that value based on information available prior to time t) the innovation.

21 Tuesday, April 2nd

21.1 Logistics

- Review Reviews due Today
- Project Draft due next week, Thursday
- Reading (Ch. 4), post by Thurs
- Quiz 6 on Thurs: First half of chapter 4.1, 4.2 (Entropy rates and markov chains). Note that 4.4 and Mixing Inequalities (and thermodynamics) will be on the next quiz (not this upcoming one).

21.2 Why cover Stochastic Processes

If our end goal is to talk about communication and channels, why divert to processes?

Answer: Stochastic Processes are a natural model for the process that generates the messages we want to send. Many messages like audio, video, text (sequence of characters with a distribution of next possible characters as opposed to future character popping out of nowhere), etc can be modeled as a process evolving over time.

This helps induce specific structural assumptions on joint dist. of possible messages, where assumptions $\Rightarrow$ results.

21.3 Entropy Rates Review

Stochastic Processes, $\mathcal{X}$ ensemble (both the space of what you can send and their respective probabilities), $\{X(t)\}_t \sim$ joint distribution.

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

where $X^{(n)} = \{X_1, X_2, \ldots, X_n\}, X_j = X(t_j)$

if there is, each time I take a step Iadd additional randomness. Thus:

Entropy increases as $n \to \infty$.

This is the sample-wise uncertainty/randomness in the trajectory (for long trajectories). This has a dual perspective if you approach the problem geometrically. As we have longer processes, the stochastic process should grow as a fn of n as n increases.

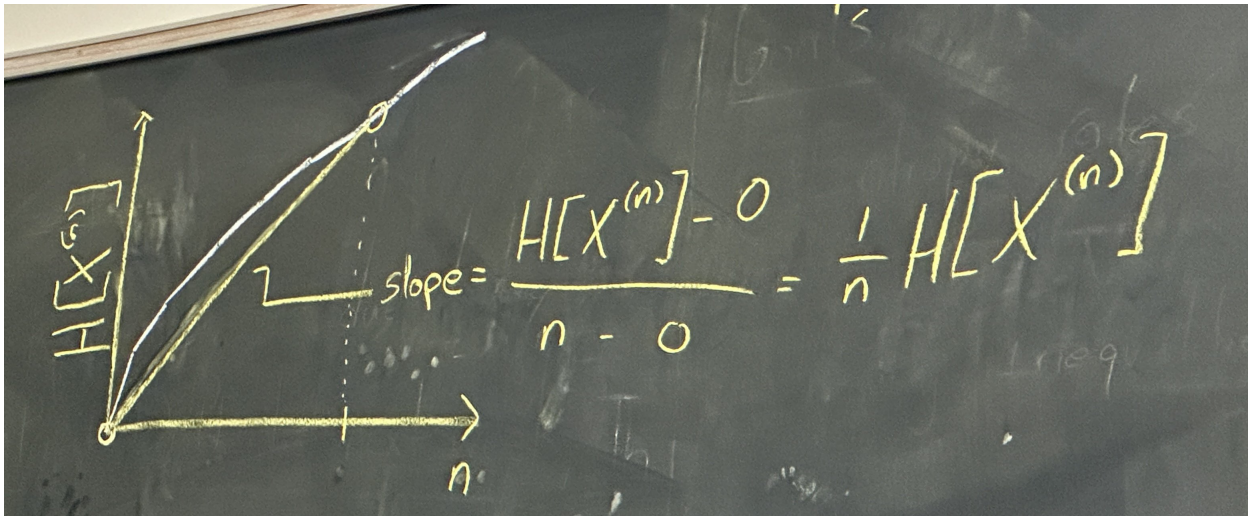


Figure 6: We examine the secant line

We use the word rate since this is the long term rate of growth in the joint entropy of a long trajectory, $H[X^{(n)}]$.

- $H[X_{n+1} \mid X^{(n)}]$
-

$$\begin{aligned} H[X^{(n+1)}] - H[X^{(n)}] &= (H[X^{(n)}] + H[X_{n+1} \mid X^{(n)}]) - H[X^{(n)}] \\ &= H[X_{n+1} \mid X^{(n)}] \end{aligned}$$

21.3.1 Alternate defn. of Entropy Rate

$$H'[\mathcal{X}] = \lim_{n \to \infty} H[X_{n+1} \mid X^{(n)}]$$

If $\mathcal{X}$ is stationary $\implies H[\mathcal{X}] = H'[\mathcal{X}]$. A stationary process means if I sample at a sequence of time, the joint distributions would not change under time (time-invariance).

21.4 Application of Entropy Rate

Recall how we studied compression earlier in the semester.

We have messages $\{X(t_j)\}_{j=1}^{n}, X^{(n)} \sim \mathcal{X}$.

We have Code: $C(X^{(n)}) \to$ string of d-ary characters.

- $(C_j \in \{\text{instantaneous codes } \mathcal{X}\})$

Note that 'dits' is the equivalent of 'bits' for d-ary character as opposed to binary.

$L(C) = \mathbb{E}_{X^{(n)}}[|c(x^{(n)})|], \quad L^*(\mathcal{X}) = \min_{c \in C}\{L(x)\} \in [H_d[X^{(n)}], H_d[X^(n)] + 1]$.

L^* should be growing as we try to encode longer messages, as we increase n.

This leads to $L_n^*(\mathcal{X}) = \frac{1}{n} L(\mathcal{X})$.

$$\lim_{n\to\infty} \frac{1}{n} L^*(\mathcal{X}) \in \lim_{n\to\infty} \frac{1}{n}(H_d[X^{(n)}] + [0,1]) \to H_d[\mathcal{X}] + \mathcal{O}\left(\frac{1}{n}\right)$$

21.5 Innovation Rates

$$\begin{aligned} \text{Entropy Rate of process } &= H[\mathcal{X}] \\ &= \text{ a minimum (best rate), the number of dits per sample needed} \\ &\quad \text{ on average to specify a trajectory.} \\ &= \text{ new info needed to specify } X_{n+1} \mid X^{(n)}(onaverage) \end{aligned}$$

Everything above is in-scope for Quiz 6.

Everything below is in-scope for Quiz 7:

21.6 Mixing Inequalities (Markov Chains)

- $\mathcal{X}$ is a (Discrete-Time) M.C., $X(0) \sim p_0, Y(0) \sim q_0, X_n \sim p_n, Y_n \sim q_n$

1. $D(p_{n+1}\|q_{n+1}) \leq D(p_n\|q_n)$, so $\{D(p_n\|q_n)\}$ is monotonically nonincreasing. 'If they follow the same rules, they should forget how they were initialized.'
   Moral:

   The process *mixes*: the cdtl. distribution of the current state gets more similar to the cdtl. distro. initialized differently.

Proof. Like most proofs in this class, if it does not involve Jensen's/convexity, we use Chain Rule:

$$D(p_{X_{n+1},X_n}\|q_{Y_{n+1},Y_n}) = \begin{cases} \text{Past:} & D(\underbrace{p_{X_n}}_{p_n}\|\underbrace{q_{Y_n}}_{q_n}) + \underbrace{D(p_{X_{n+1}|X_n}\|q_{Y_{n+1}|Y_n})}_{\text{0 since they follow the same transition probabilities}} \\ \text{Future:} & D(p_{n+1}\|q_{n+1}) + \underbrace{D(p_{X_n|X_{n+1}}\|q_{Y_n|Y_{n+1}})}_{\geq 0} \end{cases}$$

$$= \begin{cases} \text{Past:} & D(p_n\|q_n) \\ \text{Future:} & D(p_n\|q_n) - \underbrace{\dots}_{\geq 0} + \underbrace{D(p_{X_n|X_{n+1}}\|q_{Y_n|Y_{n+1}})}_{\geq 0} \end{cases}$$

$$D(p_{n+1}\|q_{n+1}) \leq D(p_n\|q_n)$$

□

21.7 What does it mean if your MC is Stationary

It means you are irreducible and aperiodic and finite. Then $\exists p_s$ which is a stationary distributions. If you initialize Y from the stationary then you stay there. Thus $D(p_{n+1}\|p_s) \leq D(p_n\|p_s)$, and $p_n \overset{n\to\infty}{\to} p_s \quad \forall p_s$.

$$D(p_n\|p_s) \overset{n\to\infty}{\to} 0, \quad D(p_n\|q_n) \overset{n\to\infty}{\to} 0 \tag{18}$$

21.8 Linkages to Data Processing

2. $I[X_1\|X_{n+1}) = I[X_{n+1}\|X_1] \leq I[X_1;X_n] = I[X_n;X_1]$, so: $\{I[X_1\|X_n]\}$ is monotonically non-increasing.
   The longer you let this process run, the less information you have about the past.
   The longer you let a MC run, the less information you have about where it started: "the process forgets where it started".
   - Then: $H[X_1 \mid X_{n+1}] \geq H[X_1 \mid X_n]$.
   - And if stationary: $H[X_{n+1} \mid X_1] \geq H[X_n \mid X_1]$. You will ultimately forget **everything** about where you started. It carries *no* information about the far past.

21.9 The second law of Thermodynamics

- $H[X_1 \mid X_{n+1}] \geq H[X_1 \mid X_n]$?
  Recall $I[A;B] = H[A] - H[A \mid B] = H[B] - H[B \mid A]$.
  DPI (Data Processing Inequality): $I[X_1;X_{n+1} \leq I[X_1;X_n]$.
  $H[X_1] - H[X_1 \mid X_{n+1}] \leq H[X_1] - H[X_1 \mid X_n] \implies -H[X_1 \mid X_{n+1}] \leq -H[X_1 \mid X_n]$
  $\implies H[X_1 \mid X_{n+1}] \geq H[X_1 \mid X_n]$. □
- $H[X_{n+1} \mid X_1] \geq H[X_n \mid X_1]$?
  DPI: $I[X_{n+1};X_1 \leq I[X_n;X_1]$.
  $H[X_{n+1}] - H[X_{n+1} \mid X_1] \leq H[X_n] - H[X_n \mid X_1]$.

If stationary by DPI:

$$\begin{aligned} H[X_{n+1}] - H[X_{n+1} \mid X_1] &\leq H[X_n] - H[X_n \mid X_1] \\ H[p_{x_{n+1}}] - H[X_{n+1}] &\leq H[p_{x_{n+1}}] - H[X_n \mid X_1] \\ H(p_s) - H[X_{n+1} \mid X_1] &\leq H(p_s) - H[X_n \mid X_1] \\ H[X_{n+1} \mid X_1] &\geq H[X_n \mid X_1] \end{aligned}$$

$H[X_{n+1} \mid X_1] \geq H[X_n \mid X_1]$ □

22 Thursday, April 4th

22.1 Logistics

- Review Reviews
- Discussion Posted
- Project Update due next Thursday
- Quiz 6 Today

22.2 Information and Thermodynamics

Consider the second law of thermodynamics, which states "entropy always increases". We want to formalize this. What exactly are we computing the entropy of? We need to take into account the random variable, distribution and the overall state of the system. Does it increase always or under certain assumptions?

The big question is if thermodynamic entropy can be related to the notion of entropy that we have developed so far.

22.3 Naive Approach

Let's try a Markov Chain approach! We'll start simple with just two states and work in discrete time. Imagine a Markov chain with two states $\boxed{1}$ and $\boxed{2}$. Let's assume transitions p_{11}, p_{12}, p_{21} and p_{22} are all nonnegative. Assume WLOG that $p_{21} >> p_{11} > 0$ and $p_{22} >> p_{12} > 0$.

Note that the steady state behaviour of this system basically sets $p^*_{22} = 1$ and $p^*_{11} = 0$ in the limit of these two assumptions. We would like to say that the entropy of some quantity is increasing in time.

Consider the entropy of $X(t)$. Clearly $H[X(t)] = 0$ in the limit since all of the probability mass is concentrated at $\boxed{2}$. However, imagine if we started the process with an initial uniform distribution?

Then we have just decreased the entropy! We have designed a process that starts with maximal entropy and reduces all the way to 0 entropy.

22.4 Thermodynamics

What is Thermo? We are not talking about your average thermoflask here. It is a statistical theory for dynamics of macroscopic observable. It studies quantities of heat, energy, temperature and pressure. Microstate refers to the detailed model. However, this is completely intractable. You want to think about this as your average Tall drink in Starbucks.

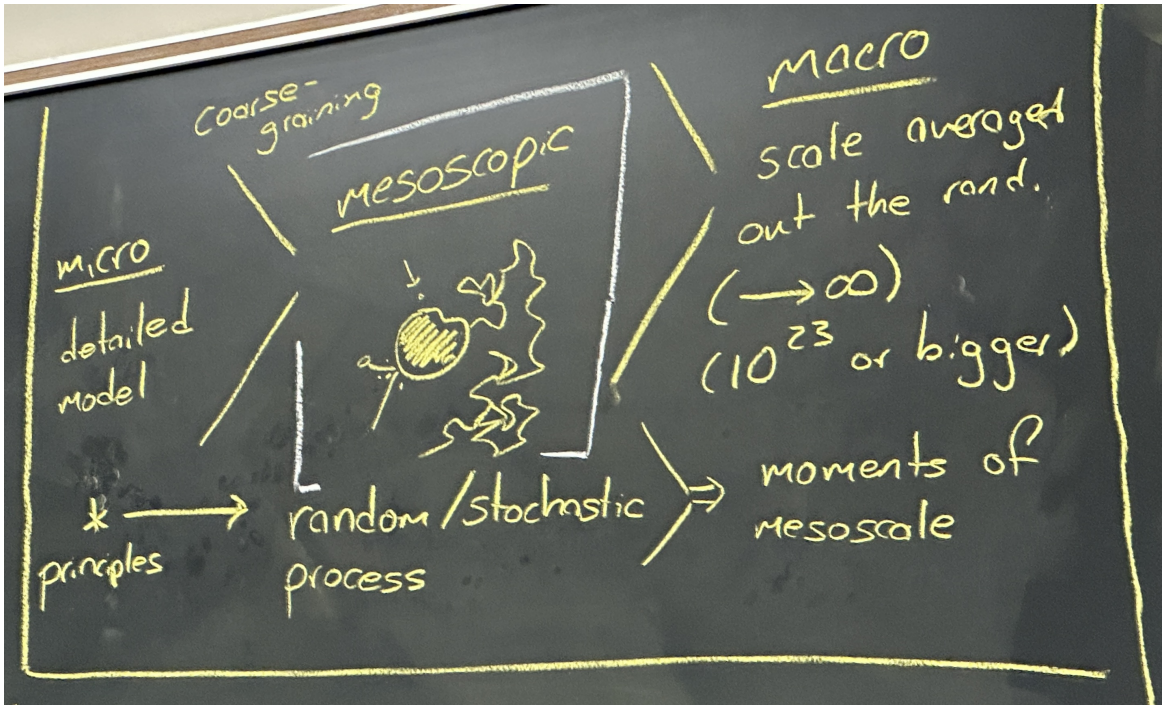
Mesoscopic states refers to "mid-sized" process. Think of this as your average Grande drink from

Starbucks.

Macroscopic state refers to the state where you average out the randomness and go to the limit of having infinite particles. Think of this as your average Venti drink from Starbucks.

22.5 Picking a System for the Mesoscopic State

22.5.1 Microscopic vs Mesoscopic vs Macroscopic



22.6 Continuous Time Markov Chains

$X(t)$ is a jump process.

- Waiting times are exponentially distributed : $T_j \sim \text{Exp}(\sum_{k\neq j} r_{kj}) = \text{Exp}(R_j)$. $R_j = \sum_{k\neq j} r_{kj}$
- Prob jump from $j \to k$ is $\frac{r_{kj}}{R_j}$
- Then: transition rate matrix: $R = \begin{cases} r_{ij} & = \text{rate} j \to i \\ r_{ii} & = -R_i = -\sum_{k\neq i} r_{k_i} \end{cases}$
- Master Equation: Let $p(t) = \text{dist} X(t), X(0) \sim p(0)$ then: $\frac{\mathrm{d}}{\mathrm{d}t}p(t) = Rp(t)$ so: $p(t) = e^{Rt}p(0)$
- Steady-State: P_s s.t. if $X(0) \sim p_s, X(t) \sim p_s$ solution to : $Rp_s = 0$
- Flux-Balance: flux $j \to i = r_{ij}p(x_j, t)$
- Complex-Balance: $\sum_{j\neq i} r_{ij}P_s(j) = R_iP_s(i)$
- Discretization: $p_n = \text{dist} X_n = X(t_n)$ if $t_n = n\Delta t$ then X_n is a discrete time MC with transition prob matrix $P(\Delta t) = e^{R\Delta t}$

What are the barebones addtl. assumptions to introduce to our stoch proc at the microscopic scale to satisfy sensible mesoscopic scale dynamics?

22.7 Conservation Laws

- total mass is conserved
- momento (momentum) is conserved
- energy is conserved – it cannot be created nor destroyed – but *why*?

22.8 Required Symmetries

1. If I knew every detail of the current state, then I would know the rule for how it changes $\implies$ dynamics only depend on the current-state.
2. The process evolves in cts.-time.
3. If we knew or could track all relevant d.o.f. then the dynamics would be *time autonomous*.
4. Time-Reversibility: Your dynamical rules (updates to your system) should be symmetric if you reverse time. This is what implies the conservation of energy.

22.9 Noether's Theorem

Every continuous symmetry of the action of a physical system with conservative forces has a corresponding conservation law.

Translational symmetry $\mapsto$ Conservation of momentum.

Rotational symmetry $\mapsto$ Conservation of angular momentum.

Time-Translational symmetry $\mapsto$ Conservation of total energy.

If your microscopic process is markov then your mesoscopic process will be Markov.

23 Thursday, April 9th

23.1 Logistics

- Reading Posted (7.1 - 7.3)
- Discussion due Thurs
- Project Part 5 due Thurs.
- Quiz 6 retake tmrw.
- Quiz 7 Thurs: Only on **Mixing Inequalities**.

23.2 Hierarchy of Models

23.2.1 Stochastic Thermodynamics

Thermodynamics is a statistical theory which works across scales of physical systems. Today we will focus on the intermediate scale of Mesoscopic. This field is known as Statistical Thermodynamics or Stochastic Thermodynamics.

Recall the conserved principles we decided to enforce from last lecture:

1. ≈Markov (assume time-scale separation/adiabatic limit allows us to assume memorylessness at this scale)
2. Time is cts.
3. Process is autonomous
4. Some(?) form of time-reversal symmetry*
    - By enforcing the correct properties, we can conserve the correct properties. Today we will get more specific to this end.
    - We could enforce a very strong form of symmetry and get results but these would not be *interesting* results as we are then restricted to a very small subset which is the class of symmetric models.

23.3 CT Review from Last Lecture

Transition matrices can be constructed from the rate matrices through the matrix exponential: $p(t+\Delta t) = e^{R\Delta t}p(0) \implies p(t) = e^{Rt}p(0)$.

23.4 Model

We have CT finite (Discrete-State) MC which is autonomous (we know or can track all relevant d.o.f. for the dynamics).

We also want reversibility, microscopic reversibility: means any time $i \to j$, you can also go $j \to i$. Thus if $r_{ij} > 0 \implies r_{ji} > 0$.

23.4.1 Stronger Symmetry

$$r_{ij} = r_{ji} \tag{19}$$

but this is *too* limiting: it only yields $p_s \sim$ Uniform.

23.4.2 A better Symmetry

Examining the stationary distribution, we want symmetry in probability flux – of steady-state ($r_{ij}p_{s_j} = r_{ji}p_{s_i} \forall i, j$): if $p = p_s$ where $p_s = \pi$ is the stationary distribution.

23.4.3 Ensembled Indistinguishability is the same as Probability flux symmetry

The former states: If $p(0) = p_s$ then it's not possible to distinguish a trajectory, run fwd in time from one run backwards in time.

Both are equivalent ways of saying that the stoch proc $X(t)$ obeys DBEs.

23.5 Probability Fluxes

$$J_{ij}(p) = r_{ij}P_j \tag{20}$$

$$\left(\frac{d}{dt}P_i(t) = \text{ flux in - flux out }\right) \tag{21}$$

23.6 Stationarity:

23.6.1 Complex Balance:

$$\underbrace{\sum_{j\neq i} r_{ij}P_{s_j}}_{\text{total in}} = \underbrace{\left[\sum_{j\neq i} r_{ji}\right] P_{s_i}}_{\text{total out}} \tag{22}$$

23.6.2 Detailed Balance:

$$r_{ij}p_{s_j} = r_{ji}p_{s_i} \forall i, j \tag{23}$$

23.7 Theorem

A CT discrete-state (finite) MC satisfies DBEs if:

$$\text{at steady-state: } J_{ij}(p_s) = J_{ji}(p_s) \tag{24}$$

$$\text{time-reversal symmetry of the ensemble if initialized at steady-state.} \tag{25}$$

$$\text{Defn.: } \mathbb{E}[\text{Production rate of a quantity } q] = \sum_{i<j} \Delta q_{ij} \left(J_{ij}(p) - J_{ji}(p)\right) \tag{26}$$

$$\text{Given } |C| \text{ cycles, you obeys DBEs } \iff \sum_{j=1}^{|C|} \ln\left(\frac{r_{x_{j+1}x_j}}{r_{x_j,x_{j+1}}}\right) = 0 \tag{27}$$

Note that all of these are 4 equivalent.

23.8 Open System

Imagine there are particles in a system (the particles can leave the system) and we want to track the current amount of particles in the finite and bounded reservoir at any given time. It is natural to utilize expectation for this task.

! What if eq. (26) is violated?

Then $\exists q$ s.t. the prod. at steady-state is > 0.

Thus at steady-state, we expect the quantity to grow but by defn. of the stationary state we know this must be true at all future times as well. If we are constantly pulling a positive amount from a bounded finite system, then we will eventually run out.

Since the stoch proc is ergodic, w.h.p., the total amnt. of quantity exchanged with the reservoir would grow w/o bound. Thus it is impossible for the reservoir to be bounded: the system is closed.

Thus at steady-state, eq. (26) $= 0$ in any q.

$$\forall \Delta q_{ij} = -\Delta q_{ji}$$

$$\sum_{j<i} \Delta q_{ij} \underbrace{\left(J_{ij}(p) - J_{ji}(p)\right)}_{=0} = 0$$

This shows that eq. (24) $\implies$ eq. (26).

23.9 Forward Trajectory

We use the notation:

$$\{X^+(t)\}_{t=0}^{T} = \{x_1, x_2, x_3, \ldots, x_n\}$$

where $x_1 = x(0), \quad x_n = x(T)$ with $\{\tau_1, \tau_2, \ldots, \tau_n\}$.

23.10 Backwards Trajectory

We use the notation:

$$X^-(s) = X^+(T - s)$$

which implies that:

$$p_s(x_1) \prod_{j=1}^{n-1} \bar{r}_{x_j} e^{-\bar{r}_{x_j}\tau_j} \frac{r_{x_{j+1},x_j}}{\bar{r}_{x_j}} = p_s(x_n) \prod_{i=1}^{n} e^{-\bar{r}_{x_i}\tau_i} \bar{r}_{x_j} \tag{28}$$

$$\implies \left(\prod_{j=1}^{n} \frac{r_{x_{j+1}x_j}}{r_{x_j x_{j+1}}} \right) \frac{p_s(x_1)}{p_s(x_n)} = 1 \tag{29}$$

If we choose $x_n = x_1$ then $\frac{p_s(x_1)}{p_s(x_n)} = 1$ no matter what the stationary probabilites specifically are. We have a cycle since we have a path which ends where it starts $\implies$ so if we take a log on both sides then we get that eq. (25) $\implies$ eq. (27).

Finally we show that eq. (25) $\implies$ eq. (24):

The simplest trajectory is one which only takes a single step. Thus we get that:

$$\frac{r_{x_2x_1} p_s(x_1)}{r_{x_1x_2} p_s(x_2)} = 1$$
$$\implies r_{ij} p_{s_j} = r_{ji} p_{s_i}$$

□

This gives us much more than a nullspace el. of the rate matrix, $Rp_s = 0$, as is given by the CBE.

23.11 A cyclic property

For any cycle, that is, a sequence of states $\{x_1, x_2, \ldots, x_n, x_{n+1} = x_1\}$, we must have:

$$\sum_{j=1}^{n} \ln \left(\frac{r_{x_{j+1}x_j}}{r_{x_1 x_{j+1}}} \right) = 0 \tag{30}$$

23.12 Line Integrals to define Energy as a quantity

If you have a path $P_1 = \{i, x_2, \ldots, x_{m-1}, j\}$ and a path $P_2 = \{i, y_2, \ldots, y_{\ell-1}, j\}$.

$$\sum_{\substack{n=1\\P_1}}^{m} \ln\left(\frac{r_{x_{n+1}x_n}}{r_{x_n x_{n+1}}}\right) = \sum_{\substack{n=1\\P_2}}^{\ell} \ln\left(\frac{r_{y_{n+1}y_n}}{r_{y_n y_{n+1}}}\right) \tag{31}$$

Then for the quantity $u(x) \in \mathbb{R}$, the value of the path $x_0 \to x_T$ sum of $\ln(\frac{r_{\rightarrow}}{r_{\leftarrow}}) = u(x_0) - u(x_T)$, $\ln\left(\frac{r_{ij}}{r_{ji}}\right) = \underbrace{u_i - u_j}_{\Delta u_{ij}}$.

Thus $\exists$ a scalar-valued quantity defined on the states: u that is conserved between system + reservoir.

This means u = energy.

23.13 Boltzmann Distribution

1. $p_s(x) \propto e^{-u(x)} \implies \frac{p_{s_i}}{p_{s_j}} = \frac{r_{ij}}{r_{ji}} = e^{-(u_j - u_i)}$.

23.14 Equipartition

2. If x, y s.t. $u(x) = u(y) \implies p_s(x) = p_s(y)$.

23.15 Relationship to Work

3. $\ln(\frac{r_{\rightarrow}}{r_{\leftarrow}}) \propto$ work required from $i \to j$.

23.16 Free Energy is decreasing

$D(p(t)\|q(t))$ is decreasing $\implies$ free energy is decr. $\implies$ properties of endothermic vs exothermic reactions.

If closed, then $u(x) = u(y) \quad \forall x, y$:

$$H[X(t)] \text{ is increasing. } \ll \text{distribution-wise 2nd law.} \tag{32}$$

24 Thursday, April 11th

24.1 Logistics

- Reading Posted
- Project Part 5 due Today
- Discussion due Today

24.2 Boltzmann

This is widely used in Physics:

$$p_s(x) \propto e^{-u(x)} \propto e^{-\frac{1}{K_B T}E(x)} \tag{33}$$

for some potential function (with an additive normalization constant) $u(x)$ and where T is temperature.

The internal energy of the system is related to the potential function $u(x)$.

24.3 Equipartition

At thermal equilibrium (this applies when the system is closed), all states with equal probabilities are equally likely.

If x, y s.t. $u(x) = u(x') \implies p_s(x) = p_s(x')$.

24.4 Relationship to Work

$W_{ij} \propto -\ln(\frac{r_{ij}}{r_{ji}})$.

We have a negative since we want low energy states to be prioritized.

24.5 Mixing

$$D(p(t)\|q_s) \searrow \tag{34}$$

$D(p(t)\|q_s)$ is mon. non-incr., converges to 0 as $t \to \infty$.

$$
\begin{aligned}
D(p(t)\|q_s) &= \mathbb{E}_{X\sim p(t)}\left[\ln\left(\frac{p(X,t)}{p_s(X)}\right)\right] \\
&= \mathbb{E}_{X\sim p(t)}\left[\ln\left(p(X,t)\right)\right] - \mathbb{E}_{X\sim p(t)}\left[\ln\left(p_s(X)\right)\right] \\
&= \mathbb{E}_{X\sim p(t)}\left[\ln\left(p(X,t)\right)\right] + \mathbb{E}_{X\sim p(t)}\left[u(x)\right] \\
&= -H[X(t)] + \frac{1}{K_B T}\mathbb{E}_{X\sim p(t)}\left[E(X)\right] \\
&= \frac{1}{K_B T}\Big(- \mathbb{E}_{x\sim p(t)}[E(X)] - K_B T \cdot H[X(t)]\Big) \\
&= \text{ Free Energy} \\
&= F[X] - F(p)
\end{aligned}
$$

24.5.1 Interpretation

$$
F = E - H \tag{35}
$$

1. We can either have $E \searrow$. We heat the reservoir (cooling the system). The system tends to release energy (as heat) which is an exothermic reaction.
2. We can also have $H \nearrow$. We absorb the hear which cools the reservoir – which is an endothermic reaction.

Entropy of a closed system $H[X(t)] \nearrow$. Thus the energy is constant no matter how you arrange the system meaning that it is impossible for the expected energy of the system to change at time t.

Thus $E(x) = E(x') \quad \forall x, x'$ means it is impossible for $\mathbb{E}[E[X(t)]] = E$.

This means if $F = E - H, F \searrow, H \nearrow$.

24.5.2 Corollary: 2nd law of thermodynamics

If the system is energetically closed, then $H[X(t)] \nearrow$.

Mathematically, if $\mathbb{E}_p[E(x)] = \mathbb{E}_{x\sim q}[E(x)] \implies E(x) = E(x') \implies u(x) = u(x')$
$\implies p_s(x) = p_s(x') \implies p_s$ is uniform.

p_s is uniform means that $\ln(\frac{r_{ij}}{r_{ji}}) = \underbrace{-(u(i) - u(j))}_{0} \implies r_{ij} = r_{ji}$.

Thus, **Free Energy is decreasing**.

24.6 Applying Method of Types

We can use the Method of Types to make it so that sample averages over the collection converge w.h.p. to $\mathbb{E}_X[q(X)]$.

24.7 Coarse-graining a Macro state

Given a reservoir which is the set of all possible temperature averages $Y = y(X)$, we can be fixed in one set of arrangements.

$$\{x \text{ s.t. } y(x) = y\}$$

24.7.1 A thermal notion of Entropy

Given $Y(t) \implies$ condtl. distro. on $X \mid Y$.

This allows us to take statements that only hold i.d. and apply them to individual trajectories.

24.8 From a distribution argument to a trajectory argument

Defn.: $y(X)$ indicates some macroscopic property of the microscopic state.

Example: Temperature may be defined as the average of the Kinetic Energy of particles in the system.

We define this s.t. $y(x) = y(x') \implies E(x) = E(x')$.

$X(t)$ evolving in time $\implies$ $Y(t) = y(X(t))$ given that $Y(t) = y \implies$ dist. $X|Y = y$ which is $\approx$Uniform over $\{x : y(x) = y\}$.

At steady state, $x \sim p_s$, what is:
$\Pr(Y = y) = \sum_{x \text{ s.t. } y(x)=y} \Pr(X = x) = (\#x \text{ s.t. } y(x) = y)e^{-\frac{1}{K_B T}E(x)}$.

But we note that this is the same as:
$= \underbrace{(\#x \text{ s.t. } y(x) = y)}_{\text{multiplicity}} \underbrace{e^{-\frac{1}{K_B T}E(y)}}_{\substack{\text{Energetic} \\ \text{state} \\ =\text{likelihood}}}$

$= e^{-\frac{1}{K_B T}\left(E(y) - K_B T \underbrace{\log((\#x \text{ s.t. } y(x) = y))}_{H[X|Y=y]}\right)}$

where $H[X \mid Y = y]$ = thermodynamic entropy: $S(y) = \ln(\#$ microstates w/ microstate $y)$.

Challenge Questions:

1. Does this imply $S(Y(t)) \nearrow$ w.h.p. if the system is closed?
2. Generalize this if the energy of the microstate is not uniquely specified by the macroscopic state. You can then rediscover the axioms that we used to define entropy w/ chain rule!

25 Thursday, April 18th

25.1 Logistics

- Discussion post by Sunday

25.2 Goals

- Channel Capacity w/ and w/o feedback
- Joint Typicality (+ AEP)

25.3 Channel Coding Theorem

If a channel $(\mathcal{X}, \mathcal{Y})$

1. Discrete: $\mathcal{X}, \mathcal{Y}$ are discrete.
2. Memoryless: Y_i given $\overset{\perp\!\!\!\perp}{} X_i$ of $y^{(i-1)}, x^{(i-1)}$.

then:

1. If $R < C$ then R is achievable.

   Recall: R is achievable if $\exists$ a set of messages $\mathcal{W}$. prior $p_{\mathcal{W}}$, and encoding $\mathcal{C}$ s.t. $|\mathcal{W}| = d^{nR}$ $|C(w)| = n$ (strings length n) and $\lambda^{(n)} = \max_{w \in cW}\{\Pr(\hat{w} \neq w \mid \mathcal{W} = w)\}$ $n \overset{\rightarrow}{\to} \infty$ 0

2. If R is achievable, $R \leq C$.

   ($R > C$ is not achievable)

where $C = \max_{P_x}\{I[X;Y]\}$.

We showed last time that $C = 1 - H[E] = 1 - H([p, 1-p])$, where E indicates an error: $X \neq Y$.

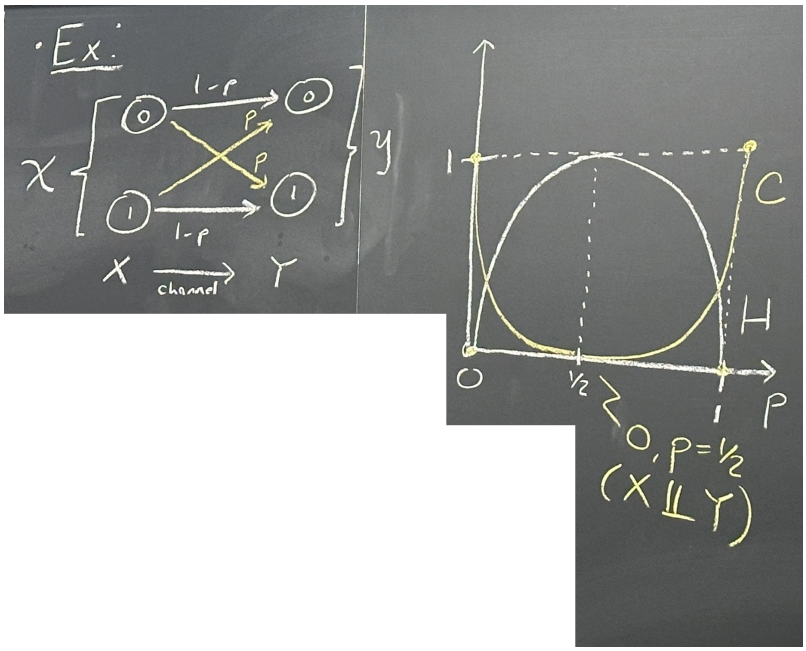
25.4 Relevant Results in this class

1. Chain Rule:

$$H[X,Y] = H[X] + H[Y \mid X]$$
$$H\left[X^{(\mu)}\right] = \sum_{i=1}^{n} H\left[X_1 \mid x^{(1-1)}\right]$$

2. Joint $\leq$ Sum: $H\left[X^{(n)}\right] \leq \sum_{i=1}^{n} H\left[X_i\right]$

3. Data Processing: $X \to Y \to W$ then $I[X;W] \leq I[X;Y]$
4. Fano's Inequality: $H[X \mid \hat{x}] \leq 1 + P_r(X \neq \hat{x}) \log(|\mathcal{X}|)$
5. Conditioning $\searrow$ Uncertainty: $H[X \mid Y] \leq H[X]$ (w/ equality if $\perp\!\!\!\perp$)



25.5 The Feedback Capacity Theorem (7.12)

A channel with feedback is illustrated in the first lecture for this class. We assume that all the received symbols are sent back immediately and noiselessly to the transmitter, which can then use them to decide which symbol to send next. Can we do better with feedback? The surprising answer is no, which we shall now prove. We define a $\left(2^{nR}, n\right)$ feedback code as a sequence of mappings $x_i\left(W, Y^{i-1}\right)$, where each x_i is a function only of the message $W \in 2^{nR}$ and the previous received values, $Y_1, Y_2, \ldots, Y_{i-1}$, and a sequence of decoding functions $g : \mathcal{Y}^n \to \left\{1, 2, \ldots, 2^{nR}\right\}$. Thus,

$$P_e^{(n)} = \Pr\left\{g\left(Y^n\right) \neq W\right\},$$

when W is uniformly distributed over $\left\{1, 2, \ldots, 2^{nR}\right\}$. Definition The capacity with feedback, C_{FB}, of a discrete memoryless channel is the supremum of all rates achievable by feedback codes.

Theorem 7.12.1 (Feedback capacity)

$$C_{\mathrm{FB}} = C = \max_{p(x)} I\left[X;Y\right].$$

Proof: Since a nonfeedback code is a special case of a feedback code, any rate that can be achieved without feedback can be achieved with feedback, and hence

$$C_{\mathrm{FB}} \geq C.$$

Proving the inequality the other way is slightly more tricky. We cannot use the same proof that we used for the converse to the coding theorem without feedback. Lemma 7.9.2 is no longer true, since X_i depends on the past received symbols, and it is no longer true that Y_i depends only on X_i and is conditionally independent of the future X 's in (7.93).

This essentially states that the Channel Coding Thm. is true w/ feedback.

25.6 Joint Typicality

Defn.: given $\left(X^{(n)}, Y^{(n)}\right)$, i.i.d from P_{x^n,Y^n} where $P_{x^n,Y^{(n)}}\left(X^{(n)}, Y^{(n)}\right) = \prod_{i=1}^{n} P_{x,y}\left(x_i, y_i\right)$

(i) $X^{(n)}$ is typical w.r.t. and $-\frac{1}{n}\log\left(p_{x^n}\left(x^{(n)}\right)\right) \in HLx$ typical w.r.t. P_y:

$$-\frac{1}{n}\log\left(p_{x^n,r^n}\left(x^{(n)}, Y^{(n)}\right)\right) \in H[X,Y] \pm \varepsilon.$$

(iii) $\underline{\text{and}}$
$\left(X^{(n)}, Y^{(n)}\right)$ is typical w.r.t. p_X

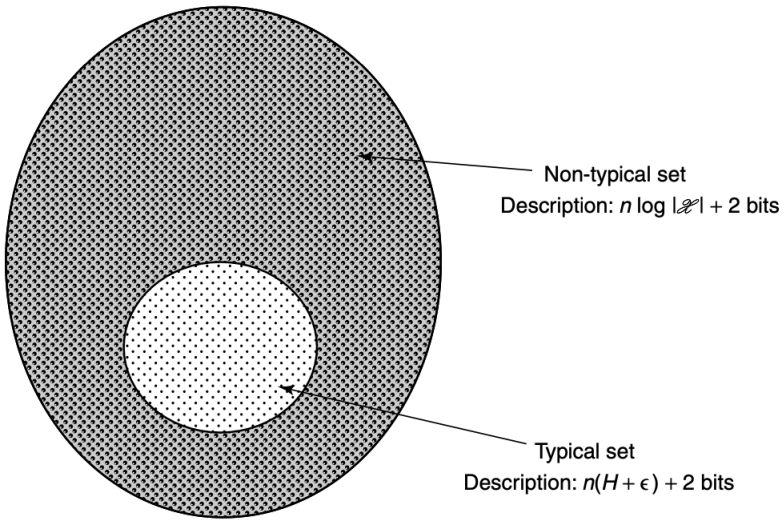


Figure 7: Source code using the typical set shows the Typical set is smaller than any Non-typical set.

25.7 Joint AEP

(i) (almost all joint seq. are typical jointly (equipartition))

$$\lim_{n\to\infty} P\left((X^n, Y^n) \in A_\varepsilon^{(n)}\right) = 1. \tag{36}$$

(ii) $P\left((X^n, Y^n) = (x^n, y^n)\right) \in\leq 2^{-n(H[X,Y]\pm\varepsilon)}, \quad (x^n, y^n) \in A_\varepsilon^n.$

(iii)

$$\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H[X,Y]+\varepsilon)} \quad \forall n \qquad \text{[size is relatively small]} \tag{37}$$

$$\geq (1-\delta)2^{n(H(X,Y)-\varepsilon)} \qquad \text{[for } n \text{ large]} \tag{38}$$

(iv) If $\tilde{X}^{(n)} \sim P_{X^n}, \tilde{Y}^{(n)} \sim P_{Y^n}, \tilde{X}^{(n)}$ is typical is $\perp\!\!\!\perp$ of $\tilde{Y}^{(n)}$ being typical, per (i) and (ii) above:

$$P\left(\left(\tilde{X}^{(n)}, \tilde{Y}^{(n)}\right) \in A_\varepsilon^{(n)}\right) \leq 2^{-n(I[X;Y]-3\varepsilon)} \qquad \forall n \tag{39}$$

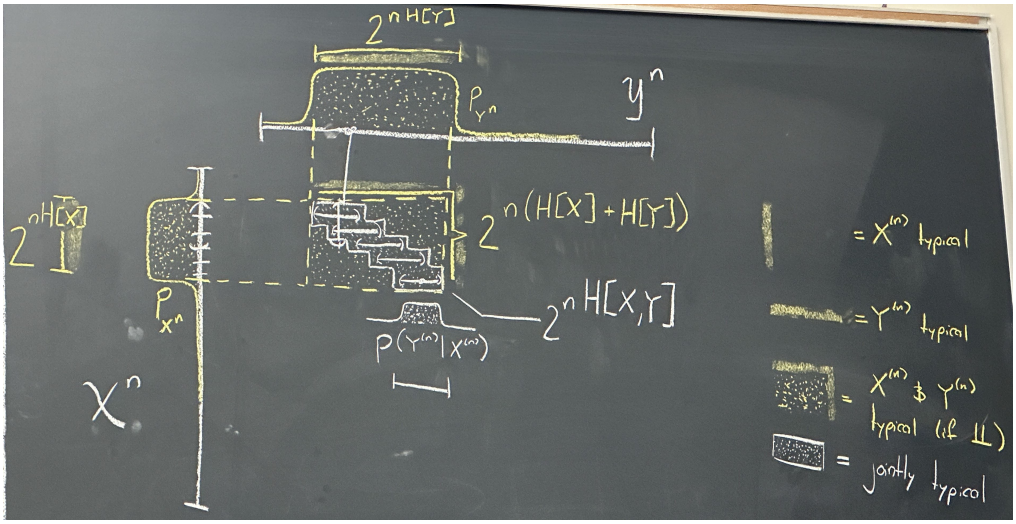
$$\geq (1-\varepsilon)2^{-n(I[X;Y]+3\varepsilon)} \qquad \text{for } n \text{ large.} \tag{40}$$

25.8 Constructing inequalities

- # messages: 2^{nR}, where $R = \frac{\log_2(|\mathcal{W}|)}{n}$
- Given $X^{(n)} = x^{(n)} : 2^{nH[Y|X]}$
- Across $2^{nH[Y]}$ typical outcomes

$$2^{nR} \cdot 2^{nH[Y|X]} \leq 2^{nH[Y]} \tag{41}$$

$$R + H[Y \mid X] \leq H[Y],\ R \leq I[X;Y] \leq C. \tag{42}$$



25.9 Proving that our packing argument is tight

2. If $R > C$, then R is <u>not</u> achievable.
If R is achievable then $R \leq C$.

- R is achievable $\implies$
- Then $P_e^{(i)} = \frac{1}{|\mathcal{W}|}\sum_w P_r(\hat{w} \neq w \mid W = w) \leq \lambda^{(n)} \longrightarrow 0$
  - Consider a model where $W \sim \text{Uniform}(\mathcal{W})$

$$P_e^{(n)} = \Pr(\hat{w} \neq w)\ n \overset{\rightarrow}{\to} \infty\ 0 \tag{43}$$

  - $\implies R \leq C$.

25.10 Putting together Past Relevant Results

$$
\begin{aligned}
nR = \log(|\mathcal{W}|) = H[W] &= H\left[W|\hat{W}\right] + I\left[W;\hat{W}\right] \\
&\leq 1 + P_e^{(n)} \log(|\mathcal{W}|) + I[W;\hat{W}] && \text{[Fano's Inequality]} \\
&= 1 + \underbrace{P_e^{(n)}}_{\to 0 \text{ as } n\to\infty} nR + I[W;\hat{W}]
\end{aligned}
$$

$$
\begin{aligned}
I\left[W; \underbrace{\hat{W}}_{g(Y^{(n)})}\right] \leq I\left[W;\hat{Y}^{(n)}\right] &= H\left[W\right] - H\left[W|Y^{(n)}\right] && \text{[Data Processing Inequality]} \\
&= I\left[\hat{Y}^{(n)};W\right] = H\left[\hat{Y}^{(n)}\right] - H\left[\hat{Y}^{(n)}|W\right]
\end{aligned}
$$

$$
\begin{aligned}
H\left[\hat{Y}^{(n)}|W\right] &= H\left[Y_1, Y_2, \ldots Y_n \mid W\right] \\
&= \sum_{i=1}^{n} H\left[Y_i^{(n)}|Y^{(i-1)}, W\right] && \text{[Chain Rule]} \\
&= \sum_{i=1}^{n} H\left[Y_i \mid X_i\right] \\
X_i &= \text{function}(Y^{(i-1)}, W) && \text{[Feedback]} \\
H[Y_i|Y^{(i-1)}, W, X_i] &= H[Y_i \mid X_i] && \text{[Memoryless]}
\end{aligned}
$$

26 Thursday, April 23rd

26.1 Logistics

- Discussion due Thurs
- Consulting OH form
- Quiz 8 on Thurs (Scope: Last unit up until today, Source Channel Separation Thm.)

26.2 Goals

- Achievability of Capacity
- AEP for ergodic sources
- Source-Channel Separation

26.3 Channel Coding Theorem

For a discrete, memoryless channel:

1. Any rate $R < C$ is achievable
2. If R is achievable, then $R \leq C$. (We showed this last time)

where:

$$C = \max_{P_x} \{I[X;Y]\}$$

Proof.
- All we need is to find a code that achieves $R, R \nearrow C$.
- Construct a random code:

1. Fix p_x
2. Sample codewords $x^{(n)}(\omega)\{X_i(\omega)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p_x$. Do this d^{nR} for a base d, you have d possible outcomes, to define the codebook.
3. Sample codewords $x^{(n)}(\omega)\{X_i(\omega)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p_x$. Do this d^{nR} for a base d, you have d possible outcomes, to define the codebook (code): C.
4. Sender & Receiver know $C, P_{Y|X}$.
5. $W \sim \text{Uniform}(\mathcal{W})$, $|\mathcal{W}| = d^{nR}$.
6. Encode $W \to X^{(n)}(\omega)$.
7. Send $X^{(n)} \overset{\to}{P_{Y|X}} Y^{(n)}$.
8. Decode $Y^{(n)} \to \hat{w} \approx \omega$.

□

26.4 Feedback Capacity Theorem

The operational capacity w/ feedback is still C = info capacity.

26.5 Joint AEP Theorem

$\left(X^{(n)}, Y^{(n)}\right) \sim P_{x^n, y^{(n)}}, \quad (X_j, Y_j)_{j=1}^n \overset{\text{i.i.d.}}{\sim} P_{x,y}$
then:

(a) $\Pr(\left(X^{(n)}, Y^{(n)}\right))$ is j.t. $\to 1$ as $n \to \infty$. (almost all sequences $\sim P_{\left(X^{(n)}, Y^{(n)}\right)}$ are j.t.
(b) If $\left(x^{(n)}, y^{(n)}\right)$ is j.t. then $\Pr(\left(X^{(n)}, Y^{(n)}\right) = \left(x^{(n)}, y^{(n)}\right)) \in d^{-n(H[X,Y] \pm \varepsilon)}$
(c)

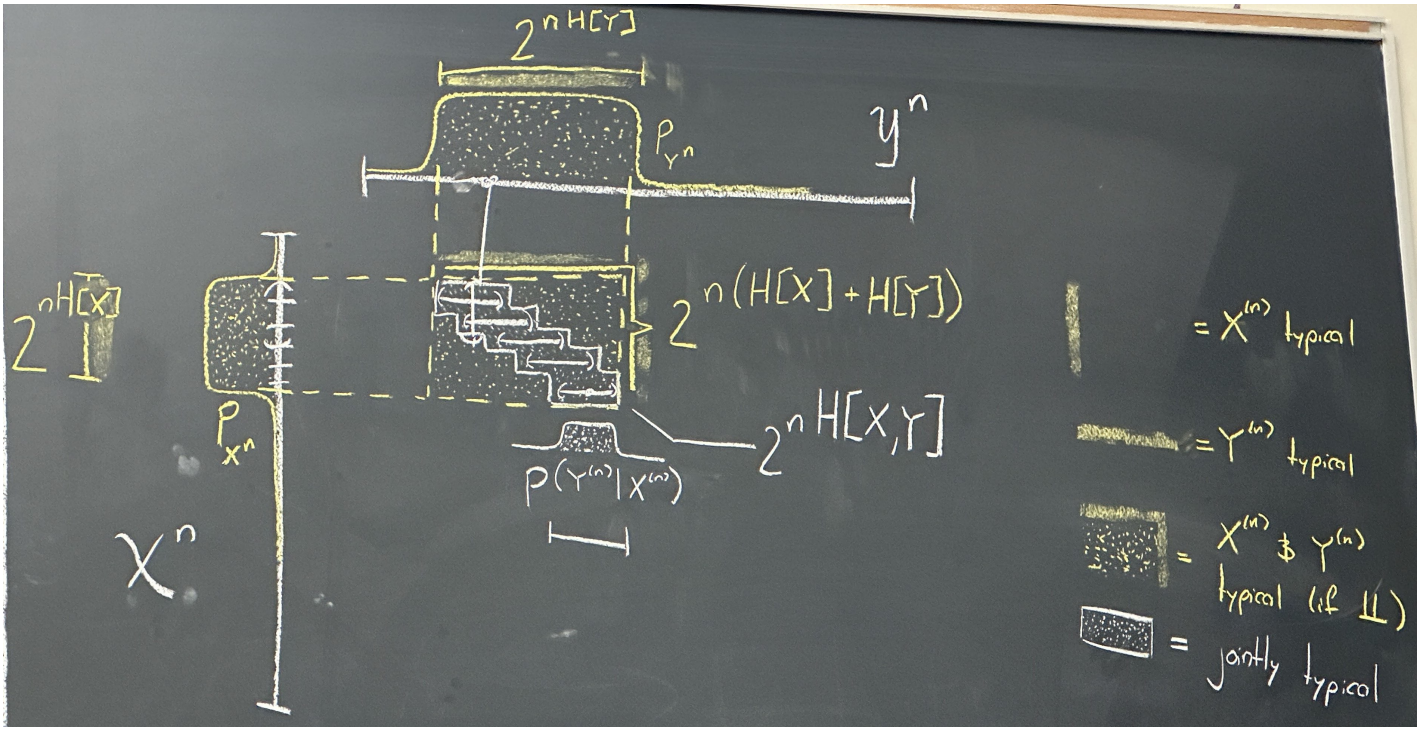
$$\begin{aligned} |\text{Jointly typical set}| &\leq d^{n(H[X,Y]+\varepsilon)} \\ &\geq (1-\varepsilon) d^{n(H[X,Y]-\varepsilon)} \qquad \text{for } n \text{ large enough} \end{aligned}$$

The set of j.t. sequences is (relatively) small.

(d) If $\tilde{X}^{(n)} \sim p_{x^n}, \tilde{Y}^{(n)} \sim p_{y^n}, \tilde{X} \perp\!\!\!\perp \tilde{Y}$ then:

$$\begin{aligned} \Pr(\left(\tilde{X}^{(n)}, \tilde{Y}^{(n)}\right) \text{ is j.t.}) &\leq d^{-n(I[X;Y]-3\varepsilon)} \\ &\geq (1-\varepsilon) d^{-n(I[X;Y]+3\varepsilon)} \qquad \text{for } n \text{ large enough} \end{aligned}$$

$\perp\!\!\!\perp$ly typical sequences are <u>rarely</u> j.t.



Here we can see that $\mathcal{X}^n$ is the input and y^n is the output, we know that we can adjust ε

26.5.1 How to decode?

Using j.t., we let $JT(\omega') = \{\left(\tilde{X}^{(n)}(\omega'), \tilde{Y}^{(n)}\right) \text{ is j.t.}\}$.

We decode by looking for an input message ω' which is j.t. with the received signal $Y^{(n)}$. If there are multiple (we cannot decode properly then we return an error).

26.5.2 When do we error?

We error in 2 cases:

1. If $JT(\omega)^c = X^{(n)}(\omega)$ is not j.t. w/ $Y^{(n)}$.
2. If $\exists \omega' \neq \omega$ s.t. $JT(\omega'), X^{(n)}(\omega')$ is also j.t. with $Y^{(n)}$.

26.5.3 Detailed Error Analysis

$$
\begin{aligned}
\Pr(\text{error}) &= \mathbb{E}_C[\Pr(\text{error} \mid C)] \\
&= \mathbb{E}_C[\mathbb{E}_W[\Pr(\text{error} \mid C, W)]] \\
&= \mathbb{E}_{C,W}[\Pr(\text{error} \mid C, W)] \\
&= \mathbb{E}_C[\Pr(\text{error} \mid C, W = 1)] \\
&= \Pr(\text{error} \mid W = 1) \\
&= \Pr(JT(1)^c \cup (\bigcup_{\omega \neq 1} JT(\omega)) \mid W = 1) \\
&\leq \underbrace{\Pr(JT(1)^c \mid W = 1)}_{\substack{Y^{(n)} \text{ is not j.t. w/ } X^{(n)}(1) \\ \leq \varepsilon}} + \sum_{\omega \neq 1} \Pr(JT(\omega') \mid W = 1) \quad \text{[Apply Union Bound]}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(JT(1)^c \mid W = 1) &\leq \varepsilon \text{ for large } n \quad \text{[by AEP]} \\
&\leq \varepsilon + \sum_{\omega' \neq 1} d^{-n(I[X;Y]-3\varepsilon)} \\
&\leq \varepsilon + \sum_{\omega' \neq 1} d^{nR} d^{-n(I[X;Y]-3\varepsilon)} \\
&\leq \varepsilon + \sum_{\omega' \neq 1} d^{-n((I[X;Y]-3\varepsilon)-R)}
\end{aligned}
$$

If $R < I[X;Y] - 3\varepsilon$ then $d^{-n(\ldots)} \searrow 0$ as $n \to \infty$. If $R < I[X;Y] - 3\varepsilon, \exists n$ large enough s.t. $\Pr(\text{error}) \leq 2\varepsilon$.

26.6 Source Model

Suppose $\mathcal{W}$ is a stochastic process.

$$W^{(n)} = \{W_i\}_{i=1}^n \sim \mathcal{W}$$

This is producing new entropy every time I draw sample, which makes us take more to encode the information.

First extend the AEP to apply to stochastic sources: Recall that the AEP is, in essence, LLN. This means we can consider ergodic (stationary) sources, where ergodic means that if we take a sample average over trajectories, it will be close to the population average.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(W_t) \overset{\text{i.p.}}{\to} \mathbb{E}_{W\sim P_s}[f(W)]$$

26.7 AEP: Shannon–McMillan–Breiman Theorem

If H is the entropy rate of a finite-valued stationary ergodic process $\{X_n\}$, then

$$-\frac{1}{n} \log p\left(X_0, \ldots, X_{n-1}\right) \to H \quad \text{with probability 1.}$$

26.8 Source-Separation Theorem

Extension to the Channel Coding Theorem for Source: if given a stochastic source which satisfies the AEP, then we can transmit messages with vanishingly probability of error.

If $V_1, V_2, \ldots V^n$ is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source-channel code with probability of error $\Pr\left(\hat{V}^n \neq V^n\right) \to 0$. Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.

26.9 What we have shown

If $R < I[X; y] - 3\varepsilon$, then for large n, average $P(\text{error}) \leq 2\varepsilon$.

1. Use $P_x = P_x^* = \underset{P_x}{\operatorname{argmax}}\{I[X;Y]\}$ then $I[X;Y] = C$ then if $R < C$, average prob of error $\to 0$ as $n \to \infty$.

2. $\Pr(\text{error}) = \mathbb{E}_C\left[P(\text{error} \mid C)\right] \leq 2\varepsilon$ then $\exists C^*$ s.t. $\Pr(\text{error} \mid C^*) \leq 2\varepsilon$.

3.

$$\Pr\left(\text{error} \mid C^*\right) \leq 2\varepsilon.$$
$$\Pr\left(\text{error} \mid C^*\right) = \mathbb{E}_{W\sim \text{ Uniformly}(\mathcal{W})}\left[\Pr\left(\text{error} \mid C^*, W\right)\right] \leq 2\varepsilon.$$

There exists a set of messages W s.t. $\Pr(\text{error} \mid C^*, W) \leq 4\varepsilon$ and there are at least $|\mathcal{W}|/2 = \frac{2^{nR}}{2} = 2^{nR-1} = 2^{n(R-1/n)}$ and discard all other messages.

Provided $\lim_{n\to\infty} R' = \lim_{n\to\infty} (R - 1/n) < C$.

27 Thursday, April 25th

27.1 Logistics

- Slip days with partners is the max
- Quiz 8 today + retakes
- Project due Monday after reading week
- Consulting OH form is now live

27.2 Wrap-Up Activity

For each of the following, discuss:

1. Something you learned that:
   (a) You found surprising, clarifying, important, useful
   (b) You would like to remember/tell a friend
2. Something you found confusing (to clarify)
3. Something you'd like to learn next

27.2.1 Foundations

27.2.2 Entropy

- Entropy as expected surprise/entropy intuition.

27.2.3 Information

- Why KL Divergence is not symmetric, taking one to be the null hypothesis (ground assumption): the order of input matters and why this is a feature and not a bug. Specifically, instead of thinking of it as measuring distance, think of how distinguishable the 2 distributions are. It is not symmetric to say how distinguishable P is from Q is the same as Q from P.
- In the past, it was just a tool to measure things and used bluntly without thought on why. This class was offered to give intuition to motivate its use in many problems.
- The Chernoff-Stein Lemma tells us that the KL Divergence is the rate of decay for how much data do you need to quantify probability of error as small enough.

27.2.4 Processes

27.2.5 Communication

- For a discrete memoryless channel Feedback doesn't help (asymptotically).
- Appreciation of Architecture.

27.3 Looking Ahead

- Kolmogorov Complexity (more of a CS topic so not focused on here, see CS 70)
- Variational Principles for Divergences (more of a STAT 210B topic)
- Application to (Kelly) Betting with log growth rates of evidence (more of a STAT 165 topic).
- Utility of perfect information: Information Markets (more of a CS 188 topic).
- Estimating Information especially in high-dimensional (intractable) space (more of a DATA C102 topic).