CO 432 Spring 2025:
Lecture Notes

Lecture notes taken, unless otherwise specified, by myself during the Spring 2025 offering of CO 432, taught by Vijay Bhattiprolu.

Lectures

Chapter 1

Introduction

Notation. I will be using my usual LaTeX typesetting conventions:

- $[n]$ means the set $\{1, 2, \dots, n\}$
- $\{0,1\}^*$ means the set of bitstrings of arbitrary length (i.e., the Kleene star)
- $\mathsf{A}, \mathsf{B}, \dots, \mathsf{Z}$ are random variables (in sans-serif)
- $\mathsf{X} = (p_1, p_2, \dots, p_k)$ means X is a discrete random variable such that $\Pr[\mathsf{X} = 1] = p_1$, $\Pr[\mathsf{X} = 2] = p_2$, etc. (abbreviate further as $\mathsf{X} = (p_i)$)

1.1 Entropy

↓ *Lecture 1 adapted from Arthur* ↓

Lecture 1
May 6

Definition 1.1.1 (entropy)
For a random variable $\mathsf{X} = (p_i)$, the entropy $H(\mathsf{X})$ is

$$H(\mathsf{X}) = -\sum_i p_i \log p_i = \sum_i p_i \log \frac{1}{p_i}.$$

Convention. By convention, we usually use $\log_2$. Also, we define entropy such that $\log_2(0) = 0$ so that impossible values do not break the formula.

Example 1.1.2. If X takes on the values a, b, c, d with probabilities 1, 0, 0, 0, respectively, then $H(\mathsf{X}) = 1 \log 1 = 0$.

If X takes on those values instead with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$, respectively, then $H(\mathsf{X}) = \frac{7}{4}$.

Fact 1.1.3. $H(\mathsf{X}) = 0$ if and only if X is a constant.

Proof. Suppose X is constant. Then, $H(\mathsf{X}) = 1 \log 1 = 0$.

Suppose $H(\mathsf{X}) = 0$. Probabilities are in $[0, 1]$, so $p_i \log \frac{1}{p_i} \geq 0$. Since $H(\mathsf{X}) = \sum_i p_i \log \frac{1}{p_i} = 0$ and each term is non-negative, each term must be zero. Thus, each p_i is either 0 or 1. We cannot have $\sum p_i > 1$, so exactly one $p_i = 1$ and the rest are zero. That is, X is constant. □

Theorem 1.1.4 (Jensen's inequality)

Let $f : \mathbb{R} \to \mathbb{R}$ be concave. That is, for any a and b in the domain of f and $\lambda \in [0, 1)$, $f(\lambda a + (1 - \lambda)b) \geq \lambda f(a) + (1 - \lambda) f(b)$. For any discrete random variable X,

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

Proof. Consider a random variable X with two values a and b, each with probabilities λ and $1 - \lambda$. Then, notice that

$$\mathbb{E}[f(\mathsf{X})] = \lambda f(a) + (1 - \lambda) f(b) \leq f(\lambda a + (1 - \lambda)b) = f(\mathbb{E}[\mathsf{X}])$$

by convexity of f.

TODO: This can be generalized by induction. □

Fact 1.1.5. Assume X is supported on $[n]$. Then, $0 \leq H(\mathsf{X}) \leq \log n$.

Proof. Start by claiming without proof that $\log n$ is concave, so we can apply Jensen's inequality.

Let $\mathsf{X}' = \frac{1}{p_i}$ with probability p_i. Then,

$$\begin{aligned} H(\mathsf{X}) &= \sum_i p_i \log \frac{1}{p_i} \\ &= \mathbb{E}\left[\log(\mathsf{X}')\right] \\ &\leq \log(\mathbb{E}[\mathsf{X}']) \\ &= \log\left(\sum p_i \frac{1}{p_i}\right) \\ &= \log n \end{aligned}$$

□

It is not a coincidence that $\log_2 n$ is the minimum number of bits to encode $[n]$.

1.2 Entropy as expected surprise

We want $S : [0, 1] \to [0, \infty)$ to capture how "surprised" we are $S(p)$ that an event with probability p happens. We want to show that under some natural assumptions, this is the only function we could have defined as entropy. In particular:

1. $S(1) = 0$, a certainty should not be surprising
2. $S(q) > S(p)$ if $p > q$, less probable should be more surprising

3. $S(p)$ is continuous in p
4. $S(pq) = S(p) + S(q)$, surprise should add for independent events. That is, if I see something twice, I should be twice as surprised.

↑ *Lecture 1 adapted from Arthur* ↑ *Lecture 2 May 8*

Proposition 1.2.1

If $S(p)$ satisfies these 4 axioms, then $S(p) = c \cdot \log_2(1/p)$ for some $c > 0$.

Proof. Suppose a function $S : [0,1] \to [0,\infty)$ exists satisfying the axioms. Let $c := S(\frac{1}{2}) > 0$.

By axiom 4 (addition), $S(\frac{1}{2^k}) = kS(\frac{1}{2})$. Likewise, $S(\frac{1}{2^{1/k}} \cdots \frac{1}{2^{1/k}}) = S(\frac{1}{2^{1/k}}) + \cdots + S(\frac{1}{2^{1/k}}) = kS(\frac{1}{2^{1/k}})$.

Then, $S(\frac{1}{2^{m/n}}) = \frac{m}{n} S(\frac{1}{2}) = \frac{m}{n} \cdot c$ for any rational m/n.

By axiom 3 (continuity), $S(\frac{1}{2^z}) = c \cdot z$ for all $z \in [0,\infty)$ because the rationals are dense in the reals. In particular, for any $p \in [0,1]$, we can write $p = \frac{1}{2^z}$ for $z = \log_2(1/p)$ and we get

$$S(p) = S\left(\frac{1}{2^z}\right) = c \cdot z = c \cdot \log_2(1/p)$$

as desired. □

We can now view entropy as expected surprise. In particular,

$$\sum_i p_i \log_2 \frac{1}{p_i} = \mathop{\mathbb{E}}_{x \sim \mathsf{X}} [S(p_x)]$$

for a random variable $\mathsf{X} = i$ with probability p_i.

1.3 Entropy as optimal lossless data compression

Suppose we are trying to compress a string consisting of n symbols drawn from some distribution.

Problem 1.3.1

What is the expected number of bits you need to store the results of n independent samples of a random variable X?

We will show this is $nH(\mathsf{X})$.

Notice that we assume that the symbols we are drawn independently, which is violated by almost all data we actually care about.

Definition 1.3.2

Let $C : \Sigma \to (\Sigma')^*$ be a code. We say C is uniquely decodable if there does not exist a collision $x, y \in \Sigma^*$, with identical encoding $C(x_1)C(x_2)\cdots C(x_k) = C(y_1)C(y_2)\cdots C(y_{k'})$.

Also, C is prefix-free (sometimes called instantaneous) if for any distinct $x, y \in \Sigma$, $C(x)$ is not a prefix of $C(y)$.

Proposition 1.3.3

Prefix-freeness is sufficient for unique decodability.

Example 1.3.4. Let $C : \{A, B, C, D\} \to \{0, 1\}^*$ where $C(A) = 11$, $C(B) = 101$, $C(C) = 100$, and $C(D) = 00$. Then, C is prefix-free and uniquely decodable.

We can easily parse 1011100001100 unambiguously as 101.11.00.00.11.00 ($BADDAD$).

Recall from CS 240 that a prefix-free code is equivalent to a trie, and we can decode it by traversing the trie in linear time.

Theorem 1.3.5 (Kraft's inequality)

A prefix-free binary code $C : \{1, \ldots, n\} \to \{0, 1\}^*$ with codeword lengths $\ell_i = |C(i)|$ exists if and only if

$$\sum_{i=1}^{n} \frac{1}{2^{\ell_i}} \leq 1.$$

Proof. Suppose $C : \{1, \ldots, n\} \to \{0, 1\}^*$ is prefix-free with codeword lengths ℓ_i. Let T be its associated binary tree and let W be a random walk on T where 0 and 1 have equal weight (stopping at either a leaf or undefined branch).

Define E_i as the event where W reaches i and $E_\varnothing$ where W falls off. Then,

$$\begin{aligned} 1 &= \Pr(E_\varnothing) + \sum_i \Pr(E_i) \\ &= \Pr(E_\varnothing) + \sum_i \frac{1}{2^{\ell_i}} && \text{(by independence)} \\ &\geq \sum_i \frac{1}{2^{\ell_i}} && \text{(probabilities are non-negative)} \end{aligned}$$

Conversely, suppose the inequality holds for some ℓ_i. WLOG, suppose $\ell_1 < \ell_2 < \cdots < \ell_n$.

Start with a complete binary tree T of depth ℓ_n. For each $i = 1, \ldots, n$, find any unassigned node in T of depth ℓ_i, delete its children, and assign it a symbol.

Now, it remains to show that this process will not fail. That is, for any loop step i, there is still some unassigned node at depth ℓ_i.

Let $P \leftarrow 2^{\ell_n}$ be the number of leaves of the complete binary tree of depth ℓ_n. After the i^{th} step, we decrease P by $2^{\ell_n - \ell_i}$. That is, after n steps,

$$\begin{aligned} P &= 2^{\ell_n} - \sum_{i=1}^{n} \frac{2^{\ell_n}}{2^{\ell_i}} \\ &= 2^{\ell_n} - 2^{\ell_n} \sum_{i=1}^{n} \frac{1}{2^{\ell_i}} \\ &\geq 0 \end{aligned}$$

by the inequality. □

Lecture 3
May 13

Recall the problem we are trying to solve:

Problem 1.3.1

What is the expected number of bits you need to store the results of n independent samples of a random variable X?

Solution (Shannon & Faro). Consider the case where X is symbol i with probability p_i. We want to encode independent samples $x_i \sim \mathsf{X}$ as $C(x_i)$ for some code $C : [n] \to \{0,1\}^*$.

Suppose for simplification that $p_i = \frac{1}{2^{\ell_i}}$ for some integers ℓ_i. Since $\sum p_i = 1$, we must have $\sum \frac{1}{2^{\ell_i}} = 1$. Then, by Kraft's inequality, there exists a prefix-free binary code $C : [n] \to \{0,1\}^*$ with codeword lengths $|C(i)| = \ell_i$. Now,

$$\mathop{\mathbb{E}}_{x_i \sim \mathsf{X}} \left[\sum_i |C(x_i)| \right] = \sum_i p_i \ell_i = \sum_i p_i \log_2 \frac{1}{p_i} = H(\mathsf{X})$$

Proceed to the general case. Suppose $\log_2 \frac{1}{p_i}$ are non-integral. Instead, use $\ell_i' = \left\lceil \log_2 \frac{1}{p_i} \right\rceil$. We still satisfy Kraft since $\sum_i \frac{1}{2^{\ell_i'}} \leq \sum_i p_i = 1$. Then,

$$\mathop{\mathbb{E}}_{x_i \sim \mathsf{X}} \left[\sum_i |C(x_i)| \right] = \sum_i p_i \ell_i' = \sum_i p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil$$

which is bounded by

$$H(\mathsf{X}) = \sum_i p_i \log_2 \frac{1}{p_i} \leq \sum_i p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil < \sum_i p_i \left(1 + \log_2 \frac{1}{p_i} \right) = H(\mathsf{X}) + 1$$

We call the code C generated by this process the Shannon–Faro code. □

We can improve on this bound $[H(\mathsf{X}), H(\mathsf{X}) + 1)$ by amortizing over longer batches of the string.

Solution (batching). For Y defined on $[n]$ equal to i with probability q_i, define the random variable $\mathsf{Y}^{(k)}$ on $[n]^k$ equal to the string $i_1 \cdots i_k$ with probability $q_{i_1} \cdots q_{i_k}$. That is, $\mathsf{Y}^{(k)}$ models k independent samples of Y.

Apply the Shannon–Fano code to $\mathsf{Y}^{(k)}$ to get an encoding of $[n]^k$ as bitstrings of expected length ℓ

satisfying $H(\mathsf{Y}^{(k)}) \le \ell \le H(\mathsf{Y}^{(k)}) + 1$.

$$\begin{aligned}
H(\mathsf{Y}^{(k)}) &= \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_1} \cdots q_{i_k}} \right] && \text{(by def'n)} \\
&= \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_1}} + \cdots + \log_2 \frac{1}{q_{i_k}} \right] && \text{(log rules)} \\
&= \sum_{j=1}^{k} \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_j}} \right] && \text{(linearity of expectation)} \\
&= \sum_{j=1}^{k} \mathop{\mathbb{E}}_{i \sim \mathsf{Y}} \left[\log_2 \frac{1}{q_i} \right] && (q_{i_j} \text{ only depends on one character}) \\
&= kH(\mathsf{Y}) && \text{(by def'n, no } j\text{-dependence in sum)}
\end{aligned}$$

For every k symbols, we use ℓ bits, i.e., $\frac{\ell}{k}$ bits per symbol. From the Shannon–Faro bound, we have

$$\begin{aligned}
\frac{H(\mathsf{Y}^{(k)})}{k} \le \frac{\ell}{k} < \frac{H(\mathsf{Y}^{(k)})}{k} + \frac{1}{k} \\
H(\mathsf{Y}) \le \frac{\ell}{k} < H(\mathsf{Y}) + \frac{1}{k}
\end{aligned}$$

Then, we have a code for Y bounded by $[H(\mathsf{Y}), H(\mathsf{Y}) + \frac{1}{k})$.

Taking a limit of some sort, we can say that we need $H(\mathsf{Y}) + o(1)$ bits. □

Definition 1.3.6 (relative entropy)

Given two discrete distributions $p = (p_i)$ and $q = (q_i)$, the relative entropy

$$D(p \parallel q) := \sum p_i \log_2 \frac{1}{q_i} - \sum_i p_i \log_2 \frac{1}{p_i} = \sum p_i \log_2 \frac{p_i}{q_i}$$

This is also known as the KL divergence.

Fact 1.3.7. $D(p \parallel q) \ge 0$ with equality exactly when $p = q$.

Proof. Define $\mathsf{X}' = \frac{p_i}{q_i}$ with probability p_i. Then,

$$D(p \parallel q) = \mathbb{E}[-\log_2 \mathsf{X}'] \ge -\log_2 E[\mathsf{X}']$$

by Jensen's inequality (as $f(x) = -\log_2 x$ is convex), and then

$$D(p \parallel q) \ge -\log_2 \sum p_i \frac{q_i}{p_i} = -\log_2 1 = 0$$ □

Proposition 1.3.8

Any prefix-free code has an expected length at least $H(\mathsf{X})$.

Proof. We can show this by interpreting the expected length $H(\mathsf{X})$ as $D(p \parallel q)$ for some q.

We will take q to be the random walk distribution corresponding to the binary tree associated to the candidate prefix-free code. □

List of Named Results

Index of Defined Terms