CS 761: Randomized Algorithms **University of Waterloo: Winter 2025**

Lecture 6

January 30, 2025

Instructor: Sepehr Assadi

Topics of this Lecture

We continue our course with a quick glance at yet another highly fundamental topic: information theory (and again, from a very limited perspective to fit into a single lecture!).

1 A Quick Intro to Information Theory

Entropy

Suppose you have a random variable X from a domain $\mathcal{X}$. How many bits to do you need to "encode" a sample X and send it over to someone else? Well, in the *worst-case*, definitely $\log|\mathcal{X}|$ bits is needed, just by pigeonhole principle. But what about *average-case* (over the randomness of X)? At this point, the answer depends on the distribution of X. For instance, if X has a uniform distribution, intuitively you cannot do much better than (almost) $\log|\mathcal{X}|$ bits even on average, while if almost all of the mass of X is on a single value x_0, you can do much better by encoding x_0 with a shorter message than the rest.

A rough idea is to do the following: start from the largest probability p of any element under X; we can have at most $1/p$ different $x \in \mathcal{X}$ with $\Pr(X = x) = p(x) = p$ as otherwise the sum of their probabilities will be more than 1; thus, we can encode each of these x's with $\log(1/p)$ bits and then move to the next value of p and continue like this. This way, for any element x, the *expected* length of encoding used for the element x is $p(x) \cdot \log(1/p(x))$ (ignoring all ceiling/floor issues, etc.).[1]

Based on this discussion, we define the **entropy** of a random variable X with PDF $p(x)$ for $x \in \mathcal{X}$ as:

$$\mathbb{H}(X) := \mathop{\mathbb{E}}_{x \sim X}\left[\log \frac{1}{p(x))}\right] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{1}{p(x)}; \tag{1}$$

(we use the convention that $0 \cdot \log(1/0) = 0$ in the definition above).

We can think of $\mathbb{H}(X)$ as a "measure" of the average length of best encoding of X.[2] To test your intuition, consider determining the entropy of each the andom variables below over the domain $\Omega = \{1, \ldots, n\}$? (compare these with the encoding length of variables you "expect" to achieve by your "own" coding scheme.)

[1] This is a very handwavy argument and is only meant to help the reader put the notion of entropy in some context; that being said, this is also not too far from what happens with Huffman coding.

[2] And again, while this is not precise, this is also not too far from the truth as Huffman coding achieves average length for a random variable X which is between $\mathbb{H}(X)$ and $\mathbb{H}(X) + 1$.

- X: uniform over $\mathcal{X}$;
- Y: deterministically equal to 1 always;
- Z: $\Pr(Z=1)=\frac{1}{2}$ and $\Pr(Z=i)=\frac{1}{2(n-1)}$ for any other $i\neq 1\in\Omega$.

Let us now establish several useful properties of entropy. We first need to recall *Jensen's inequality.*

Proposition 1 (**Jensen's inequality**). *Let f a* concave *function and X be a random variable over Ω. Then,*

$$\mathbb{E}\left[f(X)\right]\leqslant f(\mathbb{E}\left[X\right]).$$

Moreover, the equality holds iff X is deterministic or f is linear.

We now use this to establish several properties of entropy.

Property 1 (**entropy and support size**). *For a random variable X with domain $\mathcal{X}$, we have,*

$$0\leqslant\mathbb{H}(X)\leqslant\log|\mathcal{X}|.$$

The LHS is tight iff X is deterministic and the RHS is tight iff X is uniform.

Proof. The LHS is obviously true because $\mathbb{H}(X)$ is expectation of non-negative terms and the only case they are all zero is when X is deterministic. For the RHS, we apply Jensen's inequality as follows (using the fact that $\log(\cdot)$ is a concave function):

$$\mathbb{H}(X)=\mathop{\mathbb{E}}_{x\sim X}\left[\log\left(1/p(x)\right)\right]\leqslant\log\left(\mathop{\mathbb{E}}_{x\sim X}\left[1/p(x)\right]\right)=\log|\mathcal{X}|.$$

Jensen's inequality will be tight here iff all $p(x)$ values are the same (making the random variable $1/p(x)$ deterministic), thus implying the second part. □

Property 2 (**sub-additivity of entropy**). *For random variables X,Y, we have,*

$$\mathbb{H}(X,Y)\leqslant\mathbb{H}(X)+\mathbb{H}(Y);$$

moreover, the equality holds iff $X\perp Y$. Here $\mathbb{H}(X,Y)$ is simply entropy of the joint *random variable (X,Y).*

Proof. We can compute $\mathbb{H}(X,Y)-(\mathbb{H}(X)+\mathbb{H}(Y))$ as follows:

$$\begin{aligned}
\mathbb{H}(X,Y)-\mathbb{H}(X)-\mathbb{H}(Y)&=\sum_{x,y}p(x,y)\log\frac{1}{p(x,y)}-\sum_{x,y}p(x,y)\log\frac{1}{p(x)}-\sum_{x,y}p(x,y)\log\frac{1}{p(y)}\\
&\qquad\qquad(p(x)=\textstyle\sum_y p(x,y)\text{ and }p(y)=\sum_{x,y}p(x,y))\\
&=\sum_{x,y}p(x,y)\log\frac{p(x)p(y)}{p(x,y)}=\mathop{\mathbb{E}}_{(x,y)\sim(X,Y)}\left[\log\frac{p(x)p(y)}{p(x,y)}\right]\\
&\leqslant\log\left(\mathop{\mathbb{E}}_{(x,y)\sim(X,Y)}\left[\frac{p(x)p(y)}{p(x,y)}\right]\right)\qquad\text{(by Jensen's inequality)}\\
&=\log\left(\sum_{x,y}p(x)p(y)\right)=\log 1=0.
\end{aligned}$$

Moreover, Jensen's inequality above is tight whenever $p(x)p(y)=p(x,y)$ for all x,y, meaning $X\perp Y$. □

Conditional Entropy

Recall the motivating example for the entropy and now suppose that you have joint random variables (X, Y): Given a sample $(x, y) \sim (X, Y)$, how many bits do you need to encode x alone and send it to someone who already *knows* y? Again, we are interested in this question on *average* over choices of both X and Y. You can see that this question depends on the distribution of X, Y and the correlation between the two; for instance for X, Y with fixed marginals, say, both uniform, the answer would be very different when $X = Y$ so highly correlated vs. when $X \perp Y$.

This discussion brings us to the notion of **conditional entropy** defined as follows:

$$\mathbb{H}(X \mid Y) := \mathop{\mathbb{E}}_{y \sim Y} [\mathbb{H}(X \mid Y = y)] = \sum_{y \in \mathcal{Y}} p(y) \cdot \sum_{x \in \mathcal{X}} p(x \mid y) \cdot \log \frac{1}{p(x \mid y)}; \tag{2}$$

(notice that here $\mathbb{H}(X \mid Y = y)$ is just the entropy of random variable X' with distribution $X \mid Y = y$).

Examples. What are the conditional entropy of each of the random variables? (compare these with the encoding length you "expect" to achieve by your "own" coding scheme.)

- $\mathbb{H}(A \mid B)$: when A, B are independent and uniform over $\{0,1\}^n$;
- $\mathbb{H}(C \mid D)$: when C is uniform over $\{0,1\}^n$ and D is XOR of bits of D.
- $\mathbb{H}(E \mid F)$: when E is uniform over $\{0,1\}^n$ and F is the indicator random variable for $E = 0^n$.

Let us now present on some useful properties of conditional entropy.

Property 3 (**chain rule of entropy**). *For a random variables X, Y, we have,*

$$\mathbb{H}(X, Y) = \mathbb{H}(X) + \mathbb{H}(Y \mid X).$$

Proof. We can compute $\mathbb{H}(X, Y) - (\mathbb{H}(X) + \mathbb{H}(Y \mid X))$ as follows:

$$\begin{aligned}
\mathbb{H}(X, Y) - \mathbb{H}(X) - \mathbb{H}(Y \mid X) &= \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)} - \sum_{x,y} p(x, y) \log \frac{1}{p(x)} - \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{1}{p(y \mid x)} \\
&= \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)} - \sum_{x,y} p(x, y) \log \frac{1}{p(x)} - \sum_{x,y} p(x, y) \log \frac{1}{p(y \mid x)} \\
&\qquad\qquad (p(x) \cdot p(y \mid x) = p(x, y)) \\
&= \sum_{x,y} p(x, y) \log \frac{p(x) \cdot p(y \mid x)}{p(x, y)} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x, y)} \\
&= \sum_{x,y} p(x, y) \log 1 = 0.
\end{aligned}$$

□

We shall note that chain rule is one of the most important properties of entropy as it gives us *additivity*.

Another property of entropy is that conditioning can only reduce its value: after all, encoding a variable X given "extra" information Y should never become harder than if we have not been given Y in the first place. Formally,

Property 4 (**conditioning cannot increase entropy**). *For a random variables X, Y, we have,*

$$\mathbb{H}(X \mid Y) \leqslant \mathbb{H}(X);$$

moreover, the equality holds iff $X \perp Y$.

Proof. By sub-additivity $\mathbb{H}(X,Y) \leqslant \mathbb{H}(X)+\mathbb{H}(Y)$ while by chain rule $\mathbb{H}(X,Y) = \mathbb{H}(Y)+\mathbb{H}(X \mid Y)$. Plugging in these two implies the result. The second part follows from the second part of Property (2). □

Let us emphasize that this property is only true when conditioning on a random variable and not an event (i.e., realization of a random variable). Can you give an example when the latter does not hold?

Remark. This section was only a glance into the amazing area of information theory and in no ways can do justice to this field. Interested reader is referred to the excellent textbook of Cover and Thomas [CT06] as well as numerous courses on information theory tools in TCS for further background.

2 Information Theory Methods in Communication Complexity[3]

We now show one application of information theory tools: proving a lower bound on the one-way communication complexity of the Index problem studied in the previous lecture. Recall that in the Index problem, Alice is given a string $x \in \{0,1\}^n$ and Bob is given an index $i \in [n]$; the goal is for Alice to send a single message m to Bob and Bob uses m and i to output x_i. We prove the following lower bound for the Index problem (compare this with our previous lower bound from Lecture 5).

Theorem 2. *For any $\delta \in (0,1)$, any one-way δ-error protocol for Index over the uniform distribution of $x \in \{0,1\}^n$ and $i \in [n]$ requires*

$$(1 - H_2(\delta)) \cdot n,$$

communication where

$$H_2(\delta) := \delta \cdot \log \frac{1}{\delta} + (1-\delta) \cdot \log \frac{1}{(1-\delta)},$$

is the binary entropy function (i.e., entropy of a Bernoulli random variable with mean δ).[4]

To prove this theorem, we need just one more tool from information theory: *Fano's inequality*. Roughly speaking, it says that if we can use a random variable B to estimate the value of a binary random variable A, then $\mathbb{H}(A \mid B)$ should be very low (in the limit, if B deterministically fixes A, then $\mathbb{H}(A \mid B) = 0$ by Property (1)).

Property 5 (Fano's inequality). *Let A, B be two random variables and suppose that there is a function $g : supp(B) \to supp(A) = \{0,1\}$ (namely, A is a binary random variable) such that $\Pr(A \neq g(B)) = \delta$. Then, $H(A \mid B) \leqslant H_2(\delta)$.*

Proof. Define $E \in \{0,1\}$ as the indicator random variable such that $E = 1$ iff $g(B) \neq A$, i.e., the estimate of A based on $g(B)$ has an error. By the chain rule of entropy in Property (3), we have,

$$\mathbb{H}(A, E \mid B) = \mathbb{H}(E \mid B) + \mathbb{H}(A \mid E, B) = \mathbb{H}(E \mid B)$$
(where the second equality holds because A is deterministically fixed by B and E)
$$\mathbb{H}(A, E \mid B) = \mathbb{H}(A \mid B) + \mathbb{H}(E \mid A, B) = \mathbb{H}(A \mid B).$$
(where the second equality holds because E is deterministically fixed by A and B)

We thus have, using Property (4),

$$\mathbb{H}(A \mid B) = \mathbb{H}(E \mid B) \leqslant \mathbb{H}(E) \leqslant H_2(\delta),$$

where the final inequality holds because H_2 is an increasing function between 0 and 1/2 and E is a binary random variable which is 1 with probability at most δ. □

[3] The title of this section is borrowed from the title of the pioneering work of [BJKS02] which is one of the first papers that initiated the extremely fruitful approach of using information theory methods to address communication complexity questions.

[4] Note that for any constant $\delta < 1/2$, $H_2(\delta) < 1$ is a constant and thus the lower bound is $\Omega(n)$.

Proof of Theorem 2. Consider the following input distribution: we sample $x \in \{0,1\}^n$ and $i \in [n]$ independently and uniformly at random from their respective domains. Recall that as we proved in the previous lecture (the easy direction of Yao's minimax principle), we can assume without loss of generality that over this distribution, our protocol is deterministic.

Let X, M, I denote the random variable for x, m, i, respectively. Let $g(m,i)$ denote the deterministic function used by Bob to output the answer given m and i.

Since the protocol errs with probability at most δ, we know that

$$\mathop{\mathbb{E}}_{(m,i)} \Pr_x \left(x_i \neq g(m,i) \mid M = m, I = i\right) \leqslant \delta.$$

By taking A to be the random variable X_I and $B = (M, I)$ in Fano's inequality, we obtain that

$$\mathbb{H}(X_I \mid M, I) \leqslant H_2(\delta).$$

We are now going to lower bound the LHS above as follows.

$$\begin{aligned}
\mathbb{H}(X_I \mid M, I) &= \mathop{\mathbb{E}}_{i \sim I} \left[\mathbb{H}(X_I \mid M \mid I = i)\right] && \text{(by definition of conditional entropy)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{H}(X_i \mid M \mid I = i) && \text{(as } I \text{ is uniform and since } X_I = X_i \text{ conditioned on } I = i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{H}(X_i \mid M) \\
&\quad \text{(as the distribution of } X \text{ which uniquely determines both } X_i \text{ and } M \text{ is independent of the event } I = i) \\
&\geqslant \frac{1}{n} \mathbb{H}(X_1, \ldots, X_n \mid M) && \text{(by the subadditivity of entropy in Property (2))} \\
&= \frac{1}{n} \cdot \Big(\mathbb{H}(X) + \mathbb{H}(M \mid X) - \mathbb{H}(M)\Big) && \text{(by the chain rule of entropy in Property (3))} \\
&= \frac{1}{n} \cdot \Big(\mathbb{H}(X) - \mathbb{H}(M)\Big) && \text{(as } M \text{ is deterministically fixed by } X) \\
&= \frac{1}{n} \cdot \Big(n - \mathbb{H}(M)\Big).
\end{aligned}$$

This perfectly matches our intuition; to reduce the entropy of a random coordinate, we need to reduce the entropy of most coordinates; put differently, reducing the total entropy of coordinates by k bits translates to reducing the entropy of an average coordinate by only k/n bits.

Finally, by Property (1), we have,

$$\mathbb{H}(M) \leqslant \log |\mathit{supp}(M)|.$$

Plugging in this into the bound obtain above, we obtain that

$$H_2(\delta) \geqslant \frac{1}{n} \cdot \Big(n - \log |\mathit{supp}(M)|\Big) \implies \log |\mathit{supp}(M)| \geqslant (1 - H_2(\delta)) \cdot n,$$

which means that we need to have at least $2^{(1-H_2(\delta))\cdot n}$ many different messages and thus at least one message needs to be of size $(1 - H_2(\delta)) \cdot n$, proving the theorem. □

References

[BJKS02] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. Information theory methods in communication complexity. In *Proceedings of the 17th Annual IEEE Conference on Computational Complexity, Montréal, Québec, Canada, May 21-24, 2002*, pages 93–102, 2002. 4

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. 4