CS 15-859: Algorithms for Big Data Fall 2019

Lecture 9 — 2019-10-31

Prof. David Woodruff *Scribe: Wan Shen Lim*

These scribe notes cover slides 1 to 16 inclusive.

1 The INDEX Problem: Lower Bounds

To motivate this week's lecture, consider the *INDEX* problem. The *INDEX* problem is a two-player game in which Alice has a vector $x \in \{0,1\}^n$ and Bob has an index $i \in \{1, 2, \cdots, n\}$. At the end, Bob wants to output x_i with probability $\geq \frac{2}{3}$.

If both players can talk to each other, then we can solve *INDEX* with $(\log n)+1$ bits of communication:

1. Bob sends index i to Alice, which is $\log n$ bits.
2. Alice sends x_i back to Bob, which is 1 bit.
3. Bob outputs x_i.

But what if we only allow one-way communication, i.e. Alice can talk to Bob, but Bob cannot talk to Alice? Trivially, we could still solve *INDEX* with $\Omega(n)$ bits of communication – just have Alice send the n bits of x to Bob. Can we do better?

It turns out that in the one-way communication model, Alice must send at least $\Omega(n)$ bits. To prove this, we will turn to information theory.

2 Information Theory

2.1 Discrete Distribution

We will consider only discrete distributions over a finite support of size n.

A discrete distribution is a vector $p = (p_1, p_2, \cdots, p_n)$ such that

- $\forall i \in \{1, 2, \cdots, n\}\, p_i \in [0,1]$, and
- $\sum_{i=1}^{n} p_i = 1$

We say that X is a random variable with distribution p if $\mathbf{Pr}[X = i] = p_i$. Intuitively, X only takes on values from 1 to n with probabilities corresponding to p.

2.2 Entropy

Let X be a random variable with distribution p on n items, i.e. $\mathbf{Pr}[X = i] = p_i$ for $i \in \{1, \cdots, n\}$. We define the entropy of X to be

$$\mathrm{H}(X) = \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right) = \mathbb{E}\left[\log_2 \left(\frac{1}{p_i}\right)\right]$$

By convention, if $p_i = 0$ then we say $p_i \log_2 \left(\frac{1}{p_i}\right) = 0$.

Motivating entropy Entropy is a measure of uncertainty about X.

Observe that $\mathrm{H}(X) \leq \log_2(n)$, with equality achieved when $p_i = \frac{1}{n}$ for all i (a uniform distribution).

$$\sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right) = \sum_{i=1}^{n} \frac{1}{n} \log_2(n) = \frac{1}{n} \log_2(n^n) = \frac{n}{n} \log_2(n) = \log_2(n)$$

Hence informally, entropy measures how far away X is from being uniform on its domain.

Extra Wikipedia motivation The definition of entropy is motivated by Shannon's characterization of an information function I:

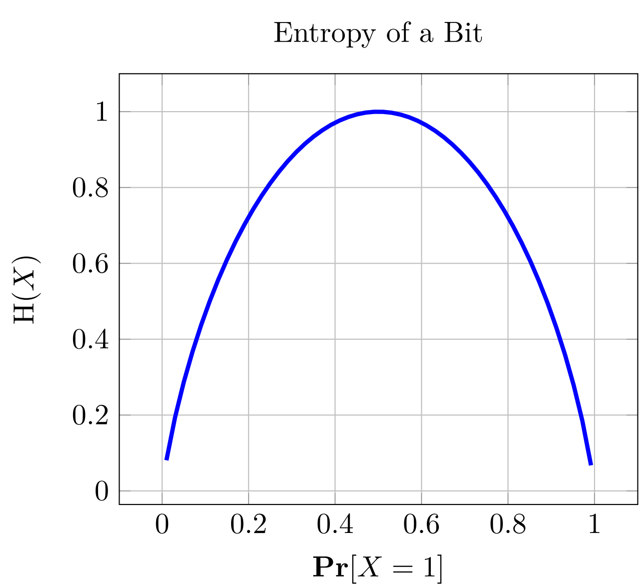
- $I(p)$ monotonically decreasing: the more likely an event, the less information it conveys
- $I(p) \geq 0$: information content is non-negative
- $I(1) = 0$: events which always occur do not convey any information
- $I(p_1 p_2) = I(p_1) + I(p_2)$: information due to independent events is additive

It turns out that $\log_2 \frac{1}{p}$ is a good function for satisfying all of the above conditions, i.e. $\log_2 \frac{1}{p}$ represents information content. But then the definition of entropy corresponds exactly to the expectation of information content, i.e. entropy can be viewed as the average information content.

Binary input There is a special case where X is a bit, i.e. a binary random variable with bias p. In this scenario, we have that

$$\mathrm{H}(X) = p \log_2 \left(\frac{1}{p}\right) + (1-p) \log_2 \left(\frac{1}{1-p}\right)$$

Notice that this is a symmetric function in p. A completely biased coin that only returns heads or that only returns tails has 0 bits of entropy, whereas a fair coin has 1 bit of entropy.



2.3 Conditional and Joint Entropy

If X and Y are random variables with distribution $(x_1, x_2, \cdots, x_m)$ and $(y_1, y_2, \cdots, y_n)$ respectively, we have

- Conditional Entropy: $\mathrm{H}(X \mid Y) = \sum_{y=1}^{n} \mathrm{H}(X \mid Y = y)\mathbf{Pr}[Y = y]$
- Joint Entropy: $\mathrm{H}(X, Y) = \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[(X, Y) = (x, y)] \log_2 \left(\frac{1}{\mathbf{Pr}[(X,y)=(x,y)]}\right)$

2.4 Chain Rule for Entropy

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

Proof.

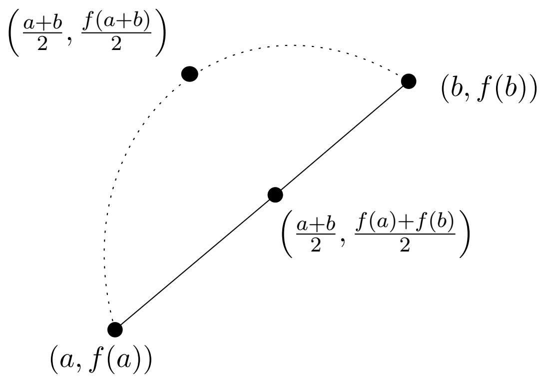
$$
\begin{aligned}
&\mathrm{H}(X, Y) \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[(X, Y) = (x, y)] \log_2 \frac{1}{\mathbf{Pr}[(X, Y) = (x, y)]} \quad \text{def of joint entropy} \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x]\mathbf{Pr}[Y = y \mid X = x] \log_2 \frac{1}{\mathbf{Pr}[X = x]\mathbf{Pr}[Y = y \mid X = x]} \quad \text{probability chain rule} \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x]\mathbf{Pr}[Y = y \mid X = x] \left(\log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) + \log_2 \left(\frac{1}{\mathbf{Pr}[Y = y \mid X = x]}\right)\right) \quad \text{log rule} \\
&= \mathbf{Pr}[X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) \left(\sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[Y = y \mid X = x]\right) \\
&\quad + \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x]\mathbf{Pr}[Y = y \mid X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[Y = y \mid X = x]}\right) \\
&= \mathbf{Pr}[X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) + \sum_{x=1}^{m} \mathbf{Pr}[X = x] \sum_{y=1}^{n} \mathbf{Pr}[Y = y \mid X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[Y = y \mid X = x]}\right) \\
&= \mathbf{Pr}[X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) + \sum_{x=1}^{m} \mathbf{Pr}[X = x] \, \mathrm{H}(Y \mid X = x) \\
&= \mathrm{H}(X) + \mathrm{H}(Y \mid X)
\end{aligned}
$$

2.5 Conditioning Cannot Increase Entropy

For any two random variables X and Y, it is true that $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$. To prove this, we rely on Jensen's inequality and facts about concave functions.

2.5.1 Jensen's Inequality and Concave Functions

Concave A function f is concave if $f\left(\frac{a+b}{2}\right) \geq \frac{f(a)+f(b)}{2}$, i.e. "the function at the average value is at least the average of the function values". An example of a concave function is $f(x) = \log(x)$.



A function is concave if it looks like the dotted line, i.e. it is always above the line drawn between any two endpoints.

Jensen's Inequality Let f be a continuous concave function and $p_1, \cdots, p_n$ be non-negative reals that sum to 1. Then for any $x_1, \cdots, x_n$, we have that $\sum_{i=1}^n p_i f(x_i) \leq f\left(\sum_{i=1}^n p_i x_i\right)$.

2.5.2 Back to conditioning not increasing entropy

$$
\begin{aligned}
&\mathrm{H}(X \mid Y) - \mathrm{H}(X) \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[Y = y]\mathbf{Pr}[X = x \mid Y = y] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x \mid Y = y]}\right) \\
&\quad - \sum_{x=1}^{m} \mathbf{Pr}[X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[Y = y]\mathbf{Pr}[X = x \mid Y = y] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x \mid Y = y]}\right) \\
&\quad - \sum_{x=1}^{m} \mathbf{Pr}[X = x] \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]}\right) \sum_{y=1}^{n} \mathbf{Pr}[Y = y \mid X = x] \quad \text{multiply by one} \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x, Y = y] \log_2 \left(\frac{\mathbf{Pr}[X = x]}{\mathbf{Pr}[X = x \mid Y = y]}\right) \quad (*) \\
&= \sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x, Y = y] \log_2 \left(\frac{\mathbf{Pr}[X = x]\mathbf{Pr}[Y = y]}{\mathbf{Pr}[(X, Y) = (x, y)]}\right) \\
&\leq \log_2 \left(\sum_{x=1}^{m} \sum_{y=1}^{n} \mathbf{Pr}[X = x, Y = y] \frac{\mathbf{Pr}[X = x]\mathbf{Pr}[Y = y]}{\mathbf{Pr}[(X, Y) = (x, y)]}\right) \quad \text{by Jensen, } \log(\cdot) \text{ concave} \\
&= \log_2(1) = 0
\end{aligned}
$$

(*) If X and Y are independent, then $\mathbf{Pr}[X = x \mid Y = y] = \mathbf{Pr}[X = x]$ and so $\log_2\left(\frac{\mathbf{Pr}[X=x]}{\mathbf{Pr}[X=x \mid Y=y]}\right) = \log_2(1) = 0$, hence we don't even need Jensen's inequality to conclude that $\mathrm{H}(X \mid Y) = \mathrm{H}(X)$.

2.6 Mutual Information

We will define the concept of mutual information.

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \mathrm{H}(Y) - \mathrm{H}(Y \mid X) = \mathrm{I}(Y;X)$$

Intuition. $\mathrm{I}(X;Y)$ refers to the information that X reveals about Y, and symmetrically also to the information that Y reveals about X.

Example. Note that $\mathrm{H}(X,X) = \sum_x \mathrm{H}(X \mid X = x)\mathbf{Pr}[X = x] = 0$. Then $\mathrm{I}(X;X) = \mathrm{H}(X) - \mathrm{H}(X \mid X) = \mathrm{H}(X)$, which matches our intuition: X just reveals itself.

It also makes sense to think about conditional mutual information. What information does X reveal about Y, given Z? We will denote this by

$$\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y,Z)$$

Note that conditioning does not necessarily give us more or less information. In particular, it is not always the case that $\mathrm{I}(X;Y \mid Z) \leq \mathrm{I}(X;Y)$ or that $\mathrm{I}(X;Y \mid Z) \geq \mathrm{I}(X;Y)$.

2.6.1 When conditioning gives us less information

Suppose we knew that $X = Y = Z$.

Then,

- $\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y,Z) = 0 - 0 = 0$
- $\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \mathrm{H}(X) - 0 = \mathrm{H}(X)$

Intuitively, Y only revealed information that Z already revealed, and we are conditioning on Z. So we learn nothing new.

2.6.2 When conditioning gives us more information

Suppose we knew that $X = Y + Z \mod 2$ for $X \sim \mathrm{Uniform}(0,1)$ and $Y \sim \mathrm{Uniform}(0,1)$.

Then, rather like the k-wise independent proofs,

- $\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y,Z) = 1 - 0 = 1$
- $\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = 1 - 1 = 0$

Intuitively, Y only revealed useful information about X after also conditioning on Z. In this case, knowing both Y and Z allow us to construct X exactly.

2.6.3 Chain Rule for Mutual Information

Claim. $\mathrm{I}(X, Y; Z) = \mathrm{I}(X; Z) + \mathrm{I}(Y; Z \mid X)$

Proof.

Using the chain rule for entropy,

$$\begin{aligned}
&\mathrm{I}(X, Y; Z) \\
&= \mathrm{H}(X, Y) - \mathrm{H}(X, Y \mid Z) \\
&= \mathrm{H}(X) + \mathrm{H}(Y \mid X) - \mathrm{H}(X \mid Z) - \mathrm{H}(Y \mid X, Z) \\
&= \mathrm{I}(X; Z) + \mathrm{I}(Y; Z \mid X)
\end{aligned}$$

And by induction,

$$\mathrm{I}(X_1, \cdots, X_n; Z) = \sum_i \mathrm{I}(X_i; Z \mid X_1, \cdots, X_{i-1})$$

2.7 Fano's Inequality

Fano's inequality is a powerful theorem for proving communication lower bounds. It is also known as the Fano converse and the Fano lemma.

2.7.1 Markov Chains

Let $X \to Y \to X'$ denote a Markov chain, i.e. X' and X are independent given Y, alternatively we can say that the past and the future are conditionally independent given the present.

We can think of X as a message being sent across a noisy channel, Y as the message received, and X' as the estimate of X that is reconstructed only from Y.

2.7.2 Data Processing Inequality

Suppose $X \to Y \to Z$ is a Markov chain. Then

$$\mathrm{I}(X; Y) \geq \mathrm{I}(X; Z)$$

Intuition. You're reconstructing Z from Y. How could Z reveal more information about X than Y? Alternatively phrased, no clever combination of the data can improve estimation.

Proof. Note that $\mathrm{I}(X; Y, Z) = \mathrm{I}(X; Z) + \mathrm{I}(X; Y \mid Z) = \mathrm{I}(X; Y) + \mathrm{I}(X; Z \mid Y)$. Hence if we show $\mathrm{I}(X; Z \mid Y) = 0$ then we are done.

But observe that $\mathrm{I}(X; Z \mid Y) = \mathrm{H}(X \mid Y) - \mathrm{H}(X \mid Y, Z)$ and that given Y, we know that X and Z are independent, and therefore $\mathrm{H}(X \mid Y, Z) = \mathrm{H}(X \mid Y)$. Hence $\mathrm{I}(X; Z \mid Y) = 0$.

Corollary. We obtain for free that $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X \mid Z)$ since

$$\mathrm{I}(X; Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) \geq \mathrm{I}(X; Z) = \mathrm{H}(X) - \mathrm{H}(X \mid Z)$$

2.7.3 Proving Fano's Inequality

Fano's inequality states that for any estimator $X' : X \to Y \to X'$ with $P_e = \mathbf{Pr}[X' \neq X]$, we have

$$\mathrm{H}(X \mid Y) \leq \mathrm{H}(P_e) + P_e \cdot \log_2 (|X| - 1)$$

Proof. Let $E = \delta(X' \neq X))$ be an indicator, i.e. $E = 1$ if $X' \neq X$ and $E = 0$ otherwise.

Then since conditioning does not increase entropy,

- $\mathrm{H}(E, X \mid X') = \mathrm{H}(X \mid X') + \mathrm{H}(E \mid X, X') = \mathrm{H}(X \mid X')$
- $\mathrm{H}(E, X \mid X') = \mathrm{H}(E \mid X') + \mathrm{H}(X \mid E, X') \leq \mathrm{H}(P_e) + \mathrm{H}(X \mid E, X')$

But then we have that

$\mathrm{H}(X \mid E, X') = \mathbf{Pr}[E = 0]\,\mathrm{H}(X \mid X', E = 0) + \mathbf{Pr}[E = 1]\,\mathrm{H}(X \mid X', E = 1) \leq (1 - P_e) \cdot 0 + P_e \log_2 (|X| - 1)$

where the last part follows because we know that E happened but $X \neq X'$ so we only have $|X| - 1$ things to consider.

Combining all of the above, we have that

$$\mathrm{H}(X \mid X') \leq \mathrm{H}(P_e) + P_e \log_2 (|X| - 1)$$

And by the data processing inequality,

$$\mathrm{H}(X \mid Y) \leq \mathrm{H}(X \mid X') \leq \mathrm{H}(P_e) + P_e \log_2 (|X| - 1)$$

2.7.4 Tightness of Fano's Inequality

Suppose the distribution p of X satisfies $p_1 \geq p_2 \geq \cdots \geq p_n$ and that Y is a constant so that $\mathrm{I}(X; Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = 0$. Then the best predictor X' of X is $X = 1$.

We have that $P_e = \mathbf{Pr}[X' \neq X] = 1 - p_1$ (the chance of predicting the first thing). By Fano's inequality, $\mathrm{H}(X \mid Y) \leq \mathrm{H}(p_1) + (1 - p_1) \log_2 (n - 1)$.

But observe that since Y is a constant, $\mathrm{H}(X) = \mathrm{H}(X \mid Y)$ and so if $p_2 = p_3 = \cdots = p_n = \frac{1-p_1}{n-1}$ then the inequality is tight. In particular, if X is drawn from $(p_1, \frac{1-p_1}{n-1}, \cdots, \frac{1-p_1}{n-1})$, then

$$\begin{aligned}
\mathrm{H}(X) &= \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right) \\
&= p_1 \log_2 \left(\frac{1}{p_1}\right) + \sum_{i=2}^{n} \frac{1 - p_1}{n - 1} \log_2 \left(\frac{n - 1}{1 - p_1}\right) \\
&= p_1 \log_2 \left(\frac{1}{p_1}\right) + (1 - p_1) \log_2 \left(\frac{1}{1 - p_1}\right) + (1 - p_1) \log_2 (n - 1) \\
&= \mathrm{H}(p_1) + (1 - p_1) \log_2 (n - 1)
\end{aligned}$$

Scribe notes end here at slide 16 inclusive.