CS 15-859: Algorithms for Big Data Fall 2020

Lecture 10 — 11/12/2020

Prof. David Woodruff *Scribe: Sylvia Zhang*

1 Motivation: Indexing

Alice has a vector $x \in \{0,1\}^n$. Bob has $i \in \{1,n\}$. INDEX$(x,i) = x_i$. Alice sends a message M to Bob based only on the vector x, and Bob computes $f(M,i) = \hat{x}_i$. If $x_i = \hat{x}_i$ with probability at least $\frac{2}{3}$, then we want to show that $|M| = \Omega(n)$. The technique used to proved such lower bounds is **information theory**.

2 Preliminaries

Consider distributions p over a finite support of size n. $p = (p_1, \cdots, p_n), p_i \in [0,1], \sum_i p_i = 1$. X is a random variable with distribution p i.e., $\mathbf{Pr}[X = i] = p_i$.

Definition (Entropy)**.** Let X be an RV with dist p on n items. The entropy $H(X) = \sum_i p_i \log_2(\frac{1}{p_i})$. If $p_i = 0$, then $p_i \log_2(\frac{1}{p_i}) = 0$ (as in the limit computation).

Entropy is a way to capture the uncertainty of a distribution. For example, a unif. dist. would have high entropy, but a constant one would have low entropy.

Claim 1. $H(X) \leq \log_2 n$. Equality holds when $p_i = \frac{1}{n}, \forall i$.If $p_i = 1$ for some i and 0 elsewhere, the entropy would be 0.

Proof. $f(x) = -x \log x$ is concave. (The average is less than the function value of the average). $\sum_{i=1}^n \frac{1}{n} f(x_i) \leq f(\frac{1}{n} \sum x_i)$. Let $x_i = p_i$. $f(x) = x \log_2 \frac{1}{x}$. $f(\frac{1}{n}) \geq \frac{1}{n} \sum (p_i \log_2 \frac{1}{p_i})$. $\frac{1}{n} \log_2 n \geq \frac{1}{n} \sum (p_i \log_2 \frac{1}{p_i})$. Hence we have proven that $H(x) \leq \log_2 n$, with the equality attained by the uniform distribution. ■

Definition (Binary Input Entropy)**.** If B is a bit with bias p, when $H(B) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$.

The entropy is maximized at $p = \frac{1}{2}$.
Let X, Y be two RVs.

Definition (Conditional entropy)**.** This is like the uncertainty of X over the choices of Y.

$$H(X \mid Y) = \sum_y H(X \mid Y = y) \mathbf{Pr}[Y = y]$$

Definition (Joint entropy)**.** This is just like the usual entropy applied to the joint distribution.

$$H(X,Y) = \sum_{x,y} \mathbf{Pr}[(X,Y) = (x,y)] \log_2 \frac{1}{\mathbf{Pr}[(X,Y) = (x,y)]}$$

Fact 1 (Chain rule for entropy)**.** $H(X,Y) = H(X) + H(Y \mid X)$

Proof.

$$H(X,Y) = \sum_{x,y} \mathbf{Pr}[(X,Y)=(x,y)] \log(\frac{1}{\mathbf{Pr}[(X,Y)=(x,y)]}) \tag{1}$$

$$= \sum_{x,y} \mathbf{Pr}[X=x]\mathbf{Pr}[Y=y \mid X=x] \log(\frac{1}{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y \mid X=x]}) \tag{2}$$

$$= \sum_{x,y} \mathbf{Pr}[X=x]\mathbf{Pr}[Y=y \mid X=x](\log(\frac{1}{\mathbf{Pr}[X=x]}) + \log(\frac{1}{\mathbf{Pr}[Y=y \mid X=x]})) \tag{3}$$

$$= H(X) + H(Y \mid X) \tag{4}$$

(1) using definition of joint entropy, (2) using definition of conditional probability, (3) splitting the log term using logarithm properties, (4) splitting the sum over x and y, using the definition of entropy. ■

3 Properties of Entropy

Fact 2 (Conditioning cannot increase entropy)**.** $H(X \mid Y) \leq H(X)$
"you know more from conditioning".

Proof. Using Jensen's inequality. Let f be a continuous concave function, let p be a distribution of support size n. For any $x_1, \cdots, x_n$

$$\sum_{i=1}^{n} p_i f(x_i) \leq f(\sum_i p_i x_i)$$

Recall that $f(x) = \log x$ is concave.

$$H(X \mid Y) - H(X) = \sum_{x,y} \mathbf{Pr}[Y=y]\mathbf{Pr}[X=x \mid Y=y] \log(\frac{1}{\mathbf{Pr}[X=x \mid Y=y]}) \tag{5}$$

$$- \sum_{x} \mathbf{Pr}[X=x] \log(\frac{1}{\mathbf{Pr}[X=x]}) \sum_{y} \mathbf{Pr}[Y=y \mid X=x] \tag{6}$$

$$= \sum_{x,y} \mathbf{Pr}[X=x, Y=y] \log \frac{\mathbf{Pr}[X=x]}{\mathbf{Pr}[X=x \mid Y=y]} \tag{7}$$

$$= \sum_{x,y} \mathbf{Pr}[X=x, Y=y] \log \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[(X,Y)=(x,y)]} \tag{8}$$

$$\leq \log(\sum_{x,y} \mathbf{Pr}[X=x, Y=y] \cdot \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[(X,Y)=(x,y)]}) \tag{9}$$

$$= \log 1 = 0 \tag{10}$$

The last term of (6) is multiplying $H(X)$ by 1, expressed in terms of conditional probabilities. (7) uses the fact that difference of log is log of ratio. (8) uses definition of conditional probability. (9) uses Jensen's inequality, with the function being the log function.If X, Y are independent, $H(X \mid Y) = H(X)$. the log goes to 1 in (9). ■

Definition (Mutual information)**.** how much info does X reveal about Y:

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

Why is it symmetric? Using the chain rule, $H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) = H(X,Y)$. If $X = Y$, the mutual information would just be the entropy of X. (X reveals everything about Y) If X and Y are independent, the mutual information would be 0. (X reveals nothing about Y).

Definition (Conditional mutual information)**.** $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$

How does $I(X;Y \mid Z)$ compare to $I(X;Y)$? Both cases could happen.

Claim 2. For certain X, Y, Z, we can have $I(X;Y \mid Z) \leq I(X;Y)$. Consider $X = Y = Z$. Recall that $I(X;Y) = H(X)$. However, if we knew Z, there's no information left for X to reveal.

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

Claim 3. For certain X, Y, Z, we can have $I(X;Y \mid Z) \geq I(X;Y)$. Consider $X = Y + Z \mod 2$, were X, Y are uniform in $\{0,1\}$. Note that due to the modular arithmetics thing, $X + Y = Z \mod 2$. Therefore, any two are p.w. independent but they aren't mutually independent.

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) = H(X) - 0 = 1 - 0 - 1$$
$$I(x;y) = h(x) - h(x \mid y) = H(X) - H(X) = 1 - 1 = 0$$

Intuitively, Y only reveals useful information about X after conditioning on Z.

Fact 3 (Chain rule for mutual info)**.** $I(X,Y;Z) = I(X;Z) + I(Y;Z \mid X)$.

It's the information that X reveals about Z plus the information that Y reveals about Z, given that we know X.

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

By induction, this generalizes to n variables:

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

Theorem 1 (Fano's Inequality)**.** *For any estimator $X' : X \to Y \to X'$ i.e., given random variable Y, X' is independent of X, if $P_e = \mathbf{Pr}[X' \neq X]$, we have*

$$H(X \mid Y) \leq H(P_e) + P_e \cdot \log(|X| - 1)$$

where $|X|$ is the support size of X.

$H(P_e)$ is the bit entropy function.
Here $X \to Y \to X'$ is a Markov Chain, meaning X' and X are independent given Y. Think of X as the past, Y as the present, and X' as the future. "Past and future are conditionally independent given the present".
In the context of Fano's inequality, X' is one's 'best guess' about X given Y.
Consider the following special case example: $X \in \{0, 1\}$. Therefore, we are only dealing with two bits going through the Y channel. $H(X \mid Y) \leq H(P_e)$. Suppose that $Y = X$, then $X' = Y = X$, so $P_e = 0$. In this case, Fano's gives us that $H(P_e) = 0$ (checks out). Suppose that Y is independent of X. In this case, the smallest possible probability error is when we output a bit uniformly independently, which is $\frac{1}{2}$. Now, $H(P_e) = 1$, and $H(X \mid Y) = 1$.
To prove Fano's inequality, we need the data processing inequality.

Theorem 2 (Data Processing Inequality)**.** *Suppose $X \to Y \to Z$ is a Markov Chain. Then $I(X;Y) \geq I(X;Z)$*
The moral takeaway is "no clever combination of the data can improve the estimation". Here, Z is the clever combination.

Proof (Data Processing).

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + I(X;Z \mid Y) \text{ (expand in two ways)}$$

Using chain rule, taking Y and Z out first respectively. Therefore, it suffices to show that $I(X;Z \mid Y) = 0$, since mutual information is nonnegative. ($H(X) \geq H(X \mid Y)$, using the lemma that conditioning only reduces entropy).

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

using the definition. But given Y, then X and Z are independent. So $H(X \mid Y,Z) = H(X \mid Y)$. A corollary is $H(X \mid Y) \leq H(X \mid Z)$, writing the original form into $H(X) - H(X \mid Y) \geq H(X) - H(X \mid Z)$). ■

Proof (Fano). Let $E = 1$ if $X' \neq X$, and $E = 0$ otherwise.

$$H(E, X \mid X') = H(X \mid X') + H(E \mid X, X') = H(X \mid X') \tag{11}$$
$$H(E, X \mid X') = H(E \mid X') + H(X \mid E, X') \leq H(P_e) + H(X \mid E, X') \tag{12}$$

(11) using chain rule, and the fact that E completely depends on X, X'. The first equality in (12) is just expanding in two ways, and the inequality is using the fact that conditioning cannot increase entropy. $H(P_e) = H(E) \geq H(E \mid X')$.

$$\begin{aligned} H(X \mid E, X') &= \mathbf{Pr}[E = 0]H(X \mid X', E = 0) + \mathbf{Pr}[E = 1]H(X \mid X', E = 1) \text{ (conditional prob)} \\ &\leq (1 - P_e) \cdot 0 + P_e \cdot \log_2(|X| - 1) \end{aligned}$$

Combining the two expressions, we get $H(X \mid X') \leq H(P_e) + P_e \cdot \log_2(|X| - 1)$ By data processing, $H(X \mid Y) \leq H(X \mid X') \leq H(P_e) + P_e \cdot \log_2(|X| - 1)$, since we are dealing with a Markov chain by assumption. ■

Statistically, Fano's inequality gives us a lower bound of the error probability. It measures how well we could predict X, in terms of the entropy of X given Y.

By the following example, we can see that Fano's inequality is actually tight: Suppose the distribution p of X satisfies $p_1 \geq p_2 \geq \cdots \geq p_n$. Suppose Y is a constant distribution. So $I(X;Y) = H(X) - H(X \mid Y) = 0$. The Best predictor X' of X is $X = 1$. $P_e = \mathbf{Pr}[X' \neq X] = 1 - p_1$. $H(X \mid Y) \leq H(p_1) + (1 - p_1)\log_2(n-1)$, as predicted by Fano's inequality. However, $H(X) = H(X \mid Y)$ and if $p_2 = p_3 = \cdots = p_n = \frac{1-p_1}{n-1}$, the inequality is tight:

$$X \sim (p_1, \frac{1-p_1}{n-1}, \cdots, \frac{1-p_1}{n-1}) \tag{13}$$

$$H(X) = \sum_i p_i \log(\frac{1}{p_i}) \tag{14}$$

$$= p_1 \log(\frac{1}{p_1}) + \sum_{i>1} \frac{1-p_1}{n-1} \log \frac{n-1}{1-p_1} \tag{15}$$

$$= p_1 \log \frac{1}{p_1} + \frac{1}{(n-1)} \sum_{i>1} (1-p_1)\log(\frac{1}{1-p_1}) + (1-p_1)\log(n-1) \tag{16}$$

$$= H(p_1) + (1-p_1)\log(n-1) \tag{17}$$

Y is a constant. So $H(X \mid Y) = H(X)$. (15), (16) uses log properties to foil the p_1 term out. Then the first two terms of (16) is the binary entropy function, and the last term is what Fano's gives. This chain of equalities shows that the equality in Fano's is attainable.