Communication Complexity 27 Sep, 2011 (@ IMSc)

9. Index function and Information theory

Lecturer: Meena Mahajan *Scribe: Fahad Panolan*

So far we were discussing the communication setting where Alice and Bob can send messages alternately to compute a function. Today we discuss one-round communication, where Alice sends a single message to Bob, and Bob computes the output from his input and the received message. A naive protocol for this problem is that Alice sends her entire input and Bob declares the output. Now the question is can we do better, especially if we allow randomness. So we want a single short message that represents Alice's input, which helps Bob to compute the function correctly with high probability. We will use information theory to obtain lower bounds for one-round communication problems.

9.1 One-round protocols and the Index Function

Definition 9.1. *A one-round protocol is a protocol where Alice sends a message to Bob, and then Bob announces the output. The one-round communication complexity of a function f, denoted by $D^1(f)$, is the cost of the best deterministic one-round protocol for f. The one-round randomized communication complexity of f, with public coin tosses and with probability of error is at most ε, is denoted $R^{1,pub}_{\varepsilon}(f)$.*

Definition 9.2. *The Index fuction is defined as follows. Alice is given $x \in \{0,1\}^n$ and Bob is given $i \in [n]$. The goal is for Bob to find x_i, i.e, the i^{th} bit in x.*

It is easy to see that $D^1(INDEX) \leq n+1$. Alice sends x and Bob outputs x_i. Now the question is: can we do better? If not, does randomization helps? Clearly, $R^{1,pub}_{1/2}(INDEX) \leq 1$: Bob just guesses a random bit r and announces the value of r. For every input x, i, the probability that $r = x_i$ is exactly $1/2$.

So to understand whether randomness helps, we should consider error bounded away from $1/2$. We will look for protocols where for every input, the probability that the protocol errs is at most $1/2 - \delta$, where $0 < \delta \leq 1/2$. It turns out that $R^{1,pub}_{1/2-\delta}(INDEX) = \Omega(\delta^2 n)$. To prove this, we need to build some background in information theory. Recall that to prove a lower bound for a randomized protocol, we introduced distributional communication complexity: the input is drawn from a distribution μ, and the protocol is deterministic, but we allow error in ε-fraction of inputs weighted by μ. We have seen the following result.

$$R^{pub}_{\varepsilon}(f) \geq \max_{\mu} D^{\mu}_{\varepsilon}(f)$$

This rsult also holds in the one-round setting. That is, if $D^{1,\mu}_{\varepsilon}(f)$ denotes the one-round communication complexity of f with respect to distribution μ and with error ε, then:

Claim 9.3. $R^{1,pub}_{\varepsilon}(f) \geq \max_{\mu} D^{1,\mu}_{\varepsilon}(f)$.

Proof. The randomized protocol is correct for every input with probability at least $1-\varepsilon$. Therefore for each μ, the randomized protocol is correct with probability at least $1-\varepsilon$. By the averaging argument, it is easy to see that there exist a random choice r such that the randomized protocol using r as the "random" string gives the correct answer for $1-\varepsilon$ fraction (weighted by μ) of inputs. □

So to prove a lower bound for one-round randomized protocols, it is enough to prove a lower bound for one-round distributional communication complexity for a suitably chosen distribution μ. We will show that $D^{1,uniform}_{1/2-\delta}(INDEX) = \Omega(\delta^2 n)$.

9.2 Information Theory

In information theory, entropy quantifies the amount of information in a message or the amount of uncertainty associated with a message. Let X be a random variable that takes value 1 with probability 1/2 and 0 with probability 1/2. Now the entropy associated with X is 1 since X is equally likely to be 1 or 0. On the other hand, if X takes value 1 with probability 1, then entropy is 0, since we can predict the value of X. What if X is neither determined nor unbiased? One way to "measure" uncertainty or information is the following setting.

Let $\Pr[X=1]=p$, and $\Pr[X=0]=1-p$. Let $Z_1, Z_2, ...Z_n$ be independent, identically distributed as X, 0-1 random variables. That is, for each i, $\Pr[Z_i=1]=p$ and $\Pr[Z_i=0]=1-p$. Let $Z=Z_1Z_2...Z_n$ be a message that Alice want to send to Bob. Clearly, n bits suffice to reveal Z. How much can Alice compress the message? Here is one encoding. Alice first sends $k=\sum_{i=1}^n Z_i$ as $\lceil \log_2 n \rceil$ bits. Then she sends an index pointing to the k-set in some pre-fixed ordering on all size k subsets of $[n]$; such an index needs $\log \binom{n}{k}$ bits.

$$\begin{aligned}
\text{Length of encoding} &= \lceil \log n \rceil + \left\lceil \log \binom{n}{k} \right\rceil \\
\mathbb{E}[\text{Encoding Length}] &= \lceil \log n \rceil + \sum_{k=0}^{n} \Pr[Z \text{ has } k \text{ 1s}] \cdot \left\lceil \log \binom{n}{k} \right\rceil \\
\mathbb{E}[\text{Fractional Encoding Length}] &= \frac{1}{n}\left(\lceil \log n \rceil + \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left\lceil \log \binom{n}{k} \right\rceil \right) \\
\lim_{n\to\infty} \mathbb{E}[\text{Fractional Encoding Length}] &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left\lceil \log \binom{n}{k} \right\rceil
\end{aligned}$$

When k is far from its expected value pn, the corresponding term above is vanishingly small. For the terms where k is close to pn, using Stirling's approximation for factorials, we can show that the above quantity converges to $p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$. Thus there is an encoding with this as the asymptotic cost per bit.

Shannon showed that aysmptotically this is the best that *any encoding* can achieve. This leads to the following definition.

Definition 9.4. *Let X be a 0-1 random variable with $\Pr[X=1]=p$. The Shannon entropy*

$H(X)$ of X is defined as

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

We often use notation $H(p, 1-p)$ or even $H(p)$ instead of $H(X)$.

This generalizes to any discrete random variable. Let X be a discrete random variable that takes N distinct values with probabilities $p_1, p_2, ..., p_N$ respectively.
Define Z_1 be a 0-1 random variable as follows

$$Z_1 = \begin{cases} 0 & \text{if } X = 1 \\ 1 & \text{otherwise} \end{cases}$$

Now, conditioned on $Z_1 = 1$ (that is, on $X \neq 1$), Z_2 is a random variable taking $N-1$ values with probabilities $\frac{p_i}{1-p_1}$ for all $i \in \{2, 3, ..., N\}$. So the entropy of X *should be* $H(X) = H(Z_1) + \Pr[Z_1 = 1] \cdot H(Z_2)$. Going with this intuition, we get

$$\begin{aligned} H(X) &= H(Z_1) + \Pr[Z_1 = 1].H(Z_2) \\ &= p_1 \log \frac{1}{p_1} + (1-p_1) \log \frac{1}{1-p_1} + (1-p_1) \sum_{i=2}^{N} \frac{p_i}{1-p_1} \log \frac{1-p_1}{p_i} \\ &= p_1 \log \frac{1}{p_1} + \left(\sum_{i=2}^{N} p_i \right) \log \frac{1}{1-p_1} + \sum_{i=2}^{N} p_i \log \frac{1-p_1}{p_i} \\ &= \sum_{i=1}^{N} p_i \log \frac{1}{p_i} \end{aligned}$$

This is indeed how the entropy is defined:

Definition 9.5. *Let X be a discrete random variable that takes N distinct values with probabilities $p_1, p_2, ..., p_N$. The entropy of X, denoted $H(X)$ or $H(p_1, \ldots, p_N)$, is defined as*

$$H(X) = \sum_{i=1}^{N} p_i \log \frac{1}{p_i}.$$

Observation 9.6. *The uniform distribution on n values (denoted U_n) has entropy $\log n$: $H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) = \log n$.*

In fact, the uniform distribution has the maximum possible entropy. To see this, we use properties of concave functions.

Definition 9.7. *A function f is concave if*

$$\forall x, y, \quad \forall 0 \leq \lambda \leq 1, \quad f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y).$$

That is, the function at the weighted average of two points is at least as large as the weighted average of the function at those points.

A function f is convex if $-f$ is concave.

Jensen's inequality allows us to extend this to averages over arbitrarily large sets.

Jensen's Inequality: For a concave function f,

$$\mathbb{E}(f(x)) \leq f(\mathbb{E}(x))$$

Claim 9.8. *If X is a discrete random variable, and support(X) is the set of values X can take with non-zero probability, then $0 \leq H(X) \leq \log |\text{support}(X)|$.*

Proof. $0 \leq H(X)$ is obvious ($\because p_i \log \frac{1}{p_i} \geq 0$ for $0 < p_i \leq 1$ and $p_i \log \frac{1}{p_i} \to 0$ as $p_i \to 0$).

Now we will prove $H(X) \leq \log |\text{support}(X)|$. Define a new random variable Z that take value p_i with probability p_i, for each i.

$$\begin{aligned}
H(X) &= \sum_{i=1}^{N} p_i \log \frac{1}{p_i} \\
&= \mathbb{E}_Z\left(\log \frac{1}{Z}\right) \\
&\leq \log\left(\mathbb{E}_Z\left(\frac{1}{Z}\right)\right) \qquad \text{(By Jensen's inequality, } \because \log \text{ is convcave)} \\
&= \log \sum_i \Pr[Z = p_i] \cdot \frac{1}{p_i} \\
&= \log(|\text{support}(X)|)
\end{aligned}$$

□

It follows that for a n-valued random variable X, $H(X) \leq \log n$. And the uniform distribution achieves this bound. In fact, this is the *only* distribution with entropy $\log n$; all other n-valued distributions have entropy strictly less than $\log n$.

We now consider the joint entropy of (possibly correlated) random variables X and Y. Following the definition of entropy, we have

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

If X and Y are independent, we expect the uncertainty in XY or the joint entropy of XY to be the sum of the individual entropies. If they are not independent, then the situation can change. So we consider conditional entropy $H(Y|X)$, that quantifies the residual uncertainty in Y even after the value of X is known. Each fixed value x for X can reduce some uncertainty in Y. The conditional entropy of Y is the residual uncertainty in Y, averaged over all values x for X.

Definition 9.9. *Let X be a random variable that takes values $x_1, x_2, ..., x_n$ with probability $p_1, p_2, ..., p_n$ respectively, and let Y be a random variable that takes values $y_1, y_2, ..., y_m$ with probability $q_1, q_2, ..., q_m$ respectively. Let $\Pr[X = x_i, Y = y_j] = r_{ij}$ for $i \in [n]$ and $j \in [m]$. The entropy of Y given X, is defined as*

$$H(Y|X) = \mathbb{E}_x\Big[H(Y|X = x)\Big]$$

Claim 9.10. $H(Y|X) = H(XY) - H(X)$.

Proof.

$$\begin{aligned}
H(Y|X) &= \mathbb{E}_x\Big[H(Y|X=x)\Big] \\
&= \sum_{i=1}^{n} p_i H\big(Y|X=x_i\big) \\
&= \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \Pr[Y=y_j|X=x_i] \log\Big(\frac{1}{\Pr[Y=y_j|X=x_i]}\Big) \\
&= \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \frac{r_{ij}}{p_i} \log\Big(\frac{p_i}{r_{ij}}\Big) \\
&= \sum_{i,j} r_{ij} \log\Big(\frac{1}{r_{ij}}\Big) - \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log\Big(\frac{1}{p_i}\Big) \\
&= \sum_{i,j} r_{ij} \log\Big(\frac{1}{r_{ij}}\Big) - \sum_{i=1}^{n} p_i \log\Big(\frac{1}{p_i}\Big) \qquad \Big(\because \sum_{j=1}^{m} r_{ij} = p_i\Big) \\
&= H(XY) - H(X)
\end{aligned}$$

□

From the above, it is straightforward to see that

Observation 9.11. *If X and Y are independent, then*

$$\begin{aligned}
H(XY) &= H(X) + H(Y) \\
H(Y|X) &= H(Y)
\end{aligned}$$

Sanity check: Can $H(Y|X)$ be more than $H(Y)$? It should not, because uncertainty in Y cannot increase if we are given more information about another variable X. Formally,

Claim 9.12. $H(Y|X) \leq H(Y)$

Proof.

$$\begin{aligned}
H(Y) - H(Y|X) &= H(X) + H(Y) - H(XY) \\
&= \sum_i p_i \log\Big(\frac{1}{p_i}\Big) + \sum_j q_j \log\Big(\frac{1}{q_j}\Big) - \sum_{i,j} r_{ij} \log\Big(\frac{1}{r_{ij}}\Big) \\
&= \sum_{i,j} r_{ij} \log\Big(\frac{r_{ij}}{p_i q_j}\Big) \qquad \Big(\because p_i = \sum_{j=1}^{m} r_{ij} \;\&\; q_j = \sum_{i=1}^{n} r_{ij}\Big) \\
&= \mathbb{E}_z\Big[\log \frac{1}{Z}\Big], \text{ where Z is a r.v with } \Pr[Z = \frac{p_i q_j}{r_{ij}}] = r_{ij} \\
&= \mathbb{E}_z[-\log Z] \\
&\geq -\log \mathbb{E}_z[Z] \qquad (\because \text{-log is convex}) \\
&\geq -\log\Big[\sum_{ij} r_{ij} \frac{p_i q_j}{r_{ij}}\Big] \\
&= -\log 1 = 0
\end{aligned}$$

□

The expression for conditional entropy gives rise to the **Entropy chain rule**:

$$H(X_1 X_2 ... X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1X_2) + ... + H(X_n|X_1X_2...X_{n-1})$$

Since conditional entropy cannot exceed unconditioned entropy, we get

Observation 9.13. $H(X_1X_2...X_n) \leq \sum_{i=1}^{n} H(X_i)$.

$H(Y|X)$ can be less than $H(Y)$, because X may carry some information about Y. We can quantify this amount of information carried as follows:

Definition 9.14. *The Mutual Information between two variables X and Y, denoted $I(X : Y)$, is defined as*

$$\begin{aligned}
I(X:Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(XY)
\end{aligned}$$

Mutual information between two variables is the reduction in the uncertainty of one variable due to the knowledge of other.

Example 9.15. *Let $Z_1, Z_2, ...Z_{10}$ be the random variables associated with tossing an unbiased coin 10 times. Let $X = Z_1...Z_7$, $Y = Z_6...Z_{10}$. Then $H(X) = 7$ and $H(Y) = 5$, while $H(X|Y) = 5$ and $H(Y|X) = 3$. Thus the mutual information between X and Y is the outcomes of the 6th and 7th tosses; $I(X : Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = 2$.*

Claim 9.16. *If $X_1, X_2, ..., X_n$ are independent, then $I(X_1X_2...X_n : Y) \geq \sum_{i=1}^{n} I(X_i : Y)$*

Proof.

$$\begin{aligned} I(X_1X_2...X_n : Y) &= H(X_1X_2....X_n) + H(Y) - H(X_1X_2...X_nY) \\ &= \Big(\sum_i H(X_i)\Big) + H(Y) - \Big(H(Y) + \sum_i H(X_i|X_1X_2...X_{i-1}Y)\Big) \\ &= \sum_i \Big(H(X_i) - H(X_i|X_1X_2...X_{i-1}Y)\Big) \\ &\geq \sum_i \Big(H(X_i) - H(X_i|Y)\Big) \\ &= \sum_i I(X_i : Y) \end{aligned}$$

□

Note: since the X_is are independent, we know that $H(X_i|X_1 \dots X_{i-1}) = H(X)$. But this does not imply that $H(X_i|X_1 \dots X_{i-1}Y) = H(X|Y)$. For instance, if X_1 and X_2 are independent and $Y = X_1 \oplus X_2$, then $H(X_1|X_2) = H(X_1) = 1$, but $H(X_1|X_2Y) = 0$ because X_2 and Y determine X_1. This is why in the proof above we can only claim an upper bound, not an equality.

Finally, we show that entropy itself is a concave function.

Claim 9.17. *Entropy is concave.*

Proof. First, let us explain what we mean by "entropy is concave". Let X and Y be any two random variables. Choose any $\lambda \in [0, 1]$, and define the random variable Z to be the outcome of the following experiment:

1. Toss a biased coin with probability of Heads being λ.
2. If the coin comes up Heads, draw a sample according to X.
3. If the coin comes up Heads, draw a sample according to Y.
4. Report whatever sample is drawn as the value of Z.

The claim is that the entropy of Z is at least $\lambda H\big[X\big] + (1-\lambda)H\big[Y\big]$.

Let B be a 0-1 random variable reporting the outcome of the biased coin's toss: Heads means $B = 0$, and Tails means $B = 1$. Then Z "copies" the value of either X or Y, depending on B. That is,

$$Z = (1-B)X + BY = \begin{cases} X & \text{if } B = 0 \\ Y & \text{if } B = 1 \end{cases}$$

$$\begin{aligned} \text{Hence} \quad H(Z) &\geq H(Z|B) && \text{because conditioning cannot increase entropy} \\ &= \mathbb{E}_b\big[H(Z|B=b)\big] && \text{definition of conditional entropy} \\ &= \lambda H\big[Z|B=0\big] + (1-\lambda)H\big[Z|B=1\big] \\ &= \lambda H\big[X\big] + (1-\lambda)H\big[Y\big] \end{aligned}$$

□