CS 229r Information Theory in Computer Science March 07, 2019

Lecture 12

Instructor: Madhu Sudan *Scribe: Alexandra Kieras*

1 Overview

1.1 Schedule

- The Set Disjointness Problem
- Information Complexity

1.2 References

The material covered in lecture today is discussed in the following:

- Bar-Yossef, Jayram, Kumar, Sivakumar

1.3 The Set Disjointness Problem

Recall the communication model discussed in previous lectures:

Alice and Bob are two players who have private values X and Y. Their goal is to compute some function f with inputs in $\{0,1\}^{2n}$; f is often (though not always) a Boolean function.

We define the set disjointness problem as the problem of determining whether two sets, X, Y, drawn from the integers $[1...n]$ are disjoint. Formally, the problem is defined as follows:

Let $X, Y \subseteq [n]$, then

$$\text{DISJ}^n(X,Y) = 1 \implies \exists_i \text{ s.t. } X_i = Y_i = 1$$

$$\text{DISJ}^n(X,Y) = 0 \text{ o/w}$$

Exercise 1. *Show that on any product distribution $\mu = \mu_x \times \mu_y$, there exists a protocol π to compute $DISJ^n(X,Y)$ with error ε and $O(\sqrt{n})$ communication complexity.*

2 Conditional Mutual Information

For (X, Y, Z) jointly distributed, we define $I(X;Y|Z)$ as the information gained about X from Y conditioned on Z.

Formally, this is defined as follows:

$$\begin{aligned} I(X;Y|Z) &= E_{Z\sim P_Z}\left[I(X|_{Z=z}, Y|_{Z=z})\right] \\ &= H(X|Z) - H(X|Y,Z) \end{aligned}$$

Note that, unlike entropy, conditioning does **not**, in general, reduce mutual information.

Example 2. *Conditioning does not always reduce mutual information.*

Suppose we have $X \perp\!\!\!\perp Y$, $Z = X \oplus Y$, for $X, Y \in Unif(\{0,1\}^n)$. Then:

$$I(X;Y) = 0$$
$$I(X;Y|Z) = n$$

Example 3. *Conditional Mutual Information of a Markov Chain*

Suppose $X - Y - Z$ is a Markov. Then, it follows that:

$$I(X;Y) \geq I(X;Y|Z)$$

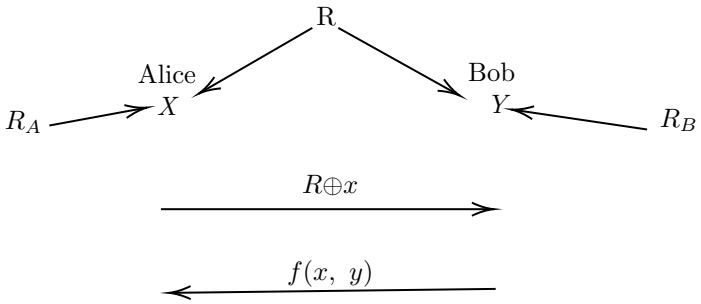
and further $I(X;Z|Y) = 0$

Exercise 4. *Prove the above claim.*

2.1 Motivation

Fix a protocol π, with error ε on all inputs while computing function f. Furthermore, fix underlying input distribution μ.

Question: How much does an observer learn about the inputs from watching the interaction?
We consider the following protocol:



Suppose the observer cannot view R (i.e, the observer cannot see the shared randomness that Alice and Bob use).

Then:

$$I(X;Y|R \oplus X, f(X,Y)) \leq H(f(X,Y))$$

That is, the observer learns little from watching the procedure, because they cannot observe the randomness R. So, we should condition on randomness R but not on R_A, R_B (the private randomness that Alice and Bob use, respectively).

2.2 Information Complexity

Definition 5. We define the *information complexity* for a protocol π over a distribution μ as:

$$IC_\mu(\pi) = I((X,Y);\pi|R)$$

The information complexity of a function f is the minimum over all protocols π which compute that function with small error.

In particular, if π is a $k-$bit protocol that $\varepsilon-$computes f - that is, if for all inputs (X, Y), π achieves at most error ε - then $IC_\mu(f) \leq k$.

3 $\Omega(n)$ Lower Bound For Disjointness

Theorem:

$$\exists \mu_n \text{ s.t. } IC_\mu(\text{DISJ}^n = \Omega(n))$$

Additional intuition:

$$IC_{\mu_n}(\text{DISJ}^n) \geq nIC_{\mu_1}(\text{DISJ}^1)$$

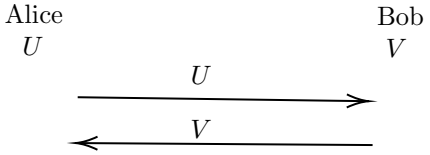
$$IC_{\mu_1}(\text{DISJ}^1) = \Omega(1)$$

While we will not directly prove this intuition, we will prove analagous results for conditional information complexity, defined in section 4.

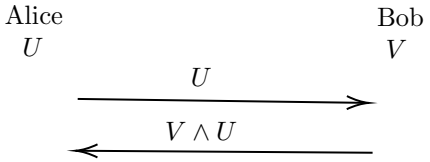
3.1 Detour: the AND function

Let $U, V \in \{0, 1\}$. Suppose that Alice has U, Bob has V, and they would like to compute $f = U \wedge V$.

Alice and Bob directly exchange U, V.



This is reasonable, and Alice and Bob will exchange two bits per function computation. However, we can do better:



In this exchange, Alice sends U to Bob, and Bob sends the computed $U \wedge V$ back.

We see that if $U = 0$, we will reveal 1 bit and if $U = 1$. we will reveal both bits. So, we will reveal $3/2$ bits in expectation.

We can come up with an even better protocol, however. Suppose Alice chooses a time $t_A \in [0, 1]$ and Bob a time $t_B \in [0, 1]$. At time t_A Alice sends U to Bob if $U = 0$. At time t_B, Bob sends 0 to Alice if $V = 0$. (The protocol ends if either Alice or Bob receives a zero.) At time $t = 1$, Bob and Alice both exchange bits only if they have not already done so.

We consider the complexity of this operation:

$$\text{if } (UV) = \left.\begin{matrix} 00 \\ 01 \\ 10 \end{matrix}\right\} \text{ we reveal 1 bit}$$

$$\text{if } (UV) = 11 \text{ we reveal both}$$

So, we reveal $\frac{5}{4}$ bits on average. (We note that this bound is not tight, because as the time approaches $t = 1$, it becomes more likely that the bit the other bit is 1.

Exercise 6. *Consider the protocol EQ, in which Alice and Bob must determine if they share the same $n-$bit string. That is:*

$$EQ(x, y) = \begin{cases} 1 & x = y \\ 0 & o/w \end{cases}$$

Prove that the communication complexity of any protocol computing this function is at least n.

3.2 Proof of DISJ bound

We return to the proof of the fact:

$$\exists \mu \text{ s.t. } IC_\mu(\text{DISJ}^n) = \Omega(n)$$

Let μ be a distribution over (X_i, Y_i) i.i.d such that:

$$(X_i, Y_i) = \begin{cases} 00 & p = \frac{1}{2} \\ 01 & p = \frac{1}{4} \\ 10 & p = \frac{1}{4} \end{cases}$$

We notice that X, Y are always disjoint! Rather than proving good performance of DISJ for this specific input, the procedure π we construct will have good performance on all inputs.

Let (X, Y, Z), with $Z \sim Unif(\{0,1\}^n)$. Then, to sample (X, Y, Z) with the distribution described above, we can execute:

For $i = 1$ to n do:

- if $Z_i = 0$ then $X_i = 0$ and $Y_i \sim Unif(\{0,1\})$
- $Z_i = 1$ then $Y_i = 0$ and $X_i \sim Unif(\{0,1\})$

3.3 Conditional Information Cost

Definition 7. We define the conditional information cost, CIC, as:

$$CIC_p(\pi) = I((X, Y); \pi | R, Z)$$

We then prove the claims:

1. $CIC_\mu(\text{DISJ}^n) \geq nCIC(\text{DISJ}^1)$
2. $CIC_\mu(\text{DISJ}^1 = \Omega(1)$

In particular, we prove Claim One in class today and Claim Two in the next lecture.

3.4 Proof of Claim One

Let M be some Markov chain $\pi \to (X,Y) \to Z$. It follows that $\pi | X, Y \perp Z | X, Y$. Following from the data processing inequality, we then have:

$$I((X,Y);\pi|R) \geq I((X,Y);\pi|R,Z)$$

Considering the right hand side of the inequality:

$$I((X,Y);\pi|R,Z) = H(X,Y|R,X) - H(X,Y|\pi,R,Z)$$

$$H(X,Y|R,Z) = \sum_{i=1}^{n} H(X_i,Y_i|R,Z,X_{<i},Y_{<i})$$

$$= \sum_{i=1}^{n} H(X_i,Y_i|Z_i)$$

$$= \sum_{i=1}^{n} H(X_i,Y_i|R,Z)$$

$$H(X,Y|\pi,R,Z) = \sum_{i=1}^{n} H(X_i,Y_i|\pi,R,Z,X_{<i},Y_{<i})$$

$$\leq \sum_{i=1}^{n} H(X_i,Y_i|\pi,R,Z)$$

So:

$$I((X,Y);\pi|R,Z) \geq \sum_{i=1}^{n} \Big(H(X_i,y_i|R,Z) - H(X_i,Y_i|\pi,R,Z) \Big)$$

$$I((X,Y);\pi|R,Z) \geq \sum_{i=1}^{n} I((X_i,Y_i);\pi|R,Z)$$

We now show that $I((X_i,Y_i);\pi,R,Z) \geq CIC(\text{DISJ}^1)$ by considering the following two communication protocols:

Protocol One:

Suppose shared random variable $w \sim Bern(0.5)$ and private randomness R_A, R_B for communication protocol between Alice and Bob. Alice computes random variable U and Bob V as follows:

- $U = 0$ if $w = 0$; otherwise it is random
- $V = 0$ id $w = 1$; otherwise it is random

Alice and Bob then compute $U \wedge V$. We see that the information revealed by this protocol is

$$I((X,Y);\pi|R,w)$$

.

Protocol Two:

We define the protocol π as follows. Let Alice and Bob share R, Z, and suppose they use Z to compute $X_1, ..., X_i$ and $Y_1, ..., Y_i$ respectively, according to μ. Suppose the output of the communication procedure is $\text{DISJ}^n(X, Y)$. The information revealed by Protocol Two is then:

$$I((X_i, Y_i); \pi | R, Z)$$

Reduction:

We show that we can compute the reduction of Protocol Two to a single input using Protocol One.

Let $X_i = U$, $Y_i = V$, and suppose that we generate Z and R using the shared randomness R of Protocol One.

We see that Protocol One will output $X_i \wedge Y_i$, which is equivalent to $\text{DISJ}^1(X_i, Y_i)$ and so $I((X, Y); \pi | R, w) = CIC(\text{DISJ}^1)$. We further see that where $j \neq i$, $X_j \wedge Y_j = 0$ by construction. Thus, the protocol will output $X_i \wedge Y_i$, which is equal to $DISJ^1$. We conclude that $I((X_i, Y_i); \pi | R, Z) \geq CIC(\text{DISJ}^1)$.

Using the equation:

$$I((X, Y); \pi | R, Z) \geq \sum_{i=1}^{n} I((X_i, Y_i); \pi | R, Z)$$

We see that:

$$I((X, Y); \pi | R, Z) \geq CIC(\text{DISJ}^1)$$
$$I((X, Y); \pi | R, Z) \geq nCIC_\mu(\text{DISJ}^1)$$

$$CIC_\mu(\text{DISJ}^n) \geq nCIC_\mu(\text{DISJ}^1)$$

Worked Solutions for Exercises:

Exercise One: Show that on any product distribution $\mu = \mu_x \times \mu_y$, with $X \perp Y$, there exists a protocol π to compute $DISJ^n(X, Y)$ with error ε and $O(\sqrt{n})$ communication complexity.

Solution:

Proof presented with guidance from article by Hemaspaandra et. al; cited fully in footnotes.

By the definition of the product distribution, we assume that sets X, Y are independently sampled such that $x_1, ..., x_n; y_1, ..., y_n \sim \mu$ are independent. We consider the restriction to the case where $|X|, |Y| = \sqrt{n}$.

From Lecture 11, we know that any $k-$bit protocol (that is, a protocol with communication complexity k) partitions the the underlying matrix M into at most 2^k monochromatic rectangles with respect to the input space.

Specifically, we consider any rectangle $R = A \times B$ created by the protocol. WLOG, we suppose that this rectangle is 0-monochromatic. Suppose that at most ε of the input pairs in $A \times B$ intersect. Equivalently, $P_{(a,b)\in R}[a \cap b = \emptyset] \geq 1 - \varepsilon$. We will show that either A or B must be small - that is, either $|A|$ or $|B| < |X| \cdot 2^{-c\sqrt{n}+1}$. We will show that if either set has a cardinality larger than this limit, the other must have a smaller cardinality. WLOG, we will show that if $|A| \geq |X| \cdot 2^{-c\sqrt{n}+1}$, then $|B| \leq |X| \cdot 2^{-c\sqrt{n}+1}$.

We consider the subset $A' \subseteq A$ such that every set $x \in A$ intersects with at most 2ε of the sets $y \in B$. Because $A \times B$ is 0-monochromatic, $|A'| \geq |A|/2 = |X|^{-c\sqrt{n}}$.

Now, we define the following process of choosing nearly-disjoint sets $X_1, ..., X_n$ from the new A': Select X_i such that the set of $X_1, ..., X_n$ are *well-separated* in that X_i contains $\frac{\sqrt{n}}{2}$ elements which are not in the previous $X_1, ..., X_{i=1}$. We lower bound the number of such possible sets that we will be able to choose by considering the sets using induction. Suppose we have chosen $i - 1$ sets and are now choosing the i^{th} set. We define:

$$Z_i = |X_i - \bigcup_{j<i} X_j|$$

If we let $Z = \bigcup_{k<i} Z_k$, we see that by the definition of the sets X_i, $|Z| = k\frac{\sqrt{n}}{3}$. If we let the total number of chosen X_i be less than $\sqrt{n}/3$, this reduces to $|z| < \frac{n}{3}$. We see that the total number of possible $x \in X$ that intersect with z in at least $\sqrt{n}2$ places is not more than:

$$\binom{n/3}{sqrtn/2}\binom{2n/3}{\sqrt{n}/2} < \binom{n}{\sqrt{n}}2^{-c\sqrt{n}}$$

and thus there must be some $x_i \in A'$ which fulfills the desired property.

Then, the full union of all of our X_i has cardinality of at least $n/3$, and each element intersects with at most 2ε elements of B. By Markov's inequality, the probability that a given $y \in G$ intersects more than $4\varepsilon k$ of the X_i is $\frac{1}{2}$. Thus, in expectation, $|G|/2$ of the $Y \in G$ intersect with more than $4\varepsilon k$ X_i. There are then $\binom{k}{4\varepsilon k}$ ways to choose this set of X_i. There are at least $n/9$ remaining X_i, giving us:

$$|G| = 2\binom{k}{4\varepsilon k}\binom{8n/9}{\sqrt{n}} < |Y| \cdot 2^{-c\sqrt{n}}$$

as desired. We thus see that no individual rectangle can have area much greater than $|X \times Y|2^{-c\sqrt{n}}2^{-c\sqrt{n}} = |X \times Y|/2^{2c\sqrt{n}}$ and thus our protocol π will create no more than $2^{c\sqrt{n}}$ rectangles, implying our communica-

tion complexity to be $\Omega(\sqrt{n})$. ■

Exercise Four: Prove the following claim:
Suppose $X - Y - Z$ is a Markov chain. Then, it follows that:

$$I(X;Y) \geq I(X;Y|Z)$$

And further, $I(X;Z|Y) = 0$.

Solution:

We consider the definition of conditional mutual information:

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

However, by the Markov property, $(X|Y) \perp Z$, and so $H(X|Y,Z)$ reduces to $H(X|Y)$. This gives the new expression:

$$I(X;Y|Z) = H(X|Z) - H(X|Y)$$

We compare to the expression for $I(X;Y)$:

$$I(X;Y) = H(X) - H(X|Y)$$

Using the property that conditioning reduces entropy, we see that $H(X|Z) \leq H(X)$. Naturally, $H(X|Y) = H(X|Y)$, and so we conclude:

$$H(X) - H(X|Y) \geq H(X|Z) - H(X|Y)$$
$$I(X;Y) \geq I(X;Y|Z) \blacksquare$$

We further consider the expression $I(X;Z|Y)$. Using the definition of conditional mutual information, we have:

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

We see again that, given Y, $X \perp Z$ and thus $H(X|Z,Y) = H(X|Y)$. Thus, our expression reduces to:

$$I(X;Z|Y) = H(X|Y) - H(X|Y) = 0 \blacksquare$$

Exercise 6:

Consider the protocol EQ, in which Alice and Bob must determine if they share the same $n-$ bit string. That is:

$$EQ(x,y) = \begin{cases} 1 & x = y \\ 0 & \text{o/w} \end{cases}$$

Prove that the communication complexity of any deterministic protocol computing this function is at least n.

Solution:

We consider *communication pairs* - that is, strings x, y that produce identical communication patterns on inputs (x, x) and (y, y). It follows that the communication pattern is identical on all four input cases $(x, x), (x, y), (y, x), (y, y)$. This follows from a brief induction: on step 1 of communication, Alice will communicate the same bit for both inputs x and y. Then, on the next step, Bob's output will also be the same because he receives the same communication from Alex and he also produces the same output on x, y, and so on.

Then, suppose towards a contradiction that there exists a protocol π for the communication such that the communication complexity of π is not more than $n-1$. This implies that Alice and Bob exchange $n-1$ bits, and so there are 2^{n-1} possible communications that could have occurred.

However, we see that there are 2^n total strings which could have comprised Alice and Bob's inputs. So, by the pigeonhole principle, there are some two inputs (x, x), (y, y) which produce the same communication pattern - and thus, the same output. But then these two input pairs also produce the same communication pattern as the pairs $(x, y), (y, x)$, and so we conclude that $\pi(x, y) = \pi(y, x) = \pi(x, x) = \pi(y, y)$. But, in particular, $EQ(x, x) = 1 \neq E(x, y)$ and so we conclude that π is not a correct deterministic protocol for computing EQ. But then there can be no correct protocol for computing EQ with computation complexity less than n. ■

Works Cited

- Arora, Sanjeev, and Boaz Barak. Computational Complexity A Modern Approach. Cambridge University Press, 2016.

- Chattopadhyay, Arkadev, and Toniann Pitassi. "The Story of Set Disjointness." ACM SIGACT News, 2010.

- Rao, Anup, and Amir Yehudayoff. Communication Complexity. 2019.