CS 229r Information Theory in Computer Science March 12, 2019

Lecture 13

Instructor: Madhu Sudan *Scribe: Tasha Schoenstein*

1 Overview

1.1 Outline

- Wrap up $\Omega(n)$ lower bound proof for DISJ
- Along the way:
  - Pinsker's Inequality
  - Hellinger Distance

1.2 Reminder

- PSET 3 due Friday (standard hard deadline, but can use late days)
- Project selection due Friday (soft deadline - good for you to do it now so that you can think about it over break)

2 Review

1. We're trying to prove lower bounds for the communication required to solve set disjointness where $DISJ^n(X,Y) = 1$ when $\exists i X_i = Y i = 1$ and is 0 otherwise.
2. In order to do this, we defined information complexity over distribution μ with error ε:

$$IC_{\mu\varepsilon}(f) = \min_{\substack{\Pi s.t. \forall X,Y \\ Pr_\Pi[\Pi(X,Y)\neq f(X,Y)]\leq\varepsilon}} [I((X,Y);\Pi|R)]$$

   where R is the public randomness of Π.
3. We then asked: which distribution is convenient to work with? Take $\mu^(n)$ on the triples (X,Y,Z) where $Z \sim \text{Unif}(\{0,1\}^n)$ and if $Z_i = 0$, $X_i = 0$ and $Y_i \sim \text{Bern}(1/2)$. If $Z_i = 1$, $Y_i = 0$ and $X_i \sim \text{Bern}(1/2)$. Note that with this distribution, the function value of disjointness is constant and 0.
4. We defined conditional information complexity

$$CIC^n_\varepsilon(\Pi) \triangleq I((X,Y);\Pi|R,Z)$$

   where Z comes from our distribution μ. This definition essentially hardwires in the distribution that we want to use (μ). We then want to show lower bounds on the communication costs using this notion because:
5. We showed last time that $IC^n_\mu(\Pi) \geq CIC^n_\varepsilon(\Pi)$.
6. We also showed that $CIC^n_\varepsilon(\Pi) \geq \sum_1^n I((X_i,Y_i),\Pi|R,Z)$. In other words, we showed that we were able to get rid of conditioning on any other X_is and Y_is.

(a) This is useful because we also showed that

$$I((X_i, Y_i), \Pi | R, Z) \geq \min_{\substack{\Pi' s.t. \\ \Pi' \text{computes} DISJ}} \{CIC^1_\varepsilon(\Pi')\}.$$

Computing disjointness is hard because we are essentially computing it coordinate by coordinate and 6a is as close as we are going to get to that.

We'll show this statement (6a) again:

$$\begin{array}{ccccc} & Alice & \leftarrow & W, R' & \rightarrow & Bob & \\ U, \tilde{X}_2, ..., \tilde{X}_n & & & & & & V, \tilde{Y}_2, ..., \tilde{Y}_n \\ X_i = \tilde{X}_i if Z = 1 & & & & & & Y_i = \tilde{Y}_i if Z = 0 \\ & Charlie & \leftarrow & Z = (W, Z_2, ...Z_n), R & \rightarrow & Dana & \\ X_1, ..., X_n, R_C & & & & & & Y_1, ..., Y_n, R_D \\ & & & \rightarrow & & & \\ & & & \leftarrow & & & \\ & & & \Pi', \Pi & & & \end{array}$$

The output of $\Pi(X, Y)$ is hopefully $DISJ^n(X, Y)$ and the output of $\Pi'(U, V)$ is hopefully $DISJ(U, V) = U \wedge V$. Notice that Charlie and Dana don't use Z. We need this because we want to say that $I((X, Y); \Pi | R, Z)$ is small so that we can design a protocol where Alice and Bob get one bit each and want to compute the AND of their bits. If we want the mutual information and the outputs to match of the two communications, we let X_1 be U, Y_1 be V, $R' = RZ_2...Z_n$, and Alice and Bob get private randomness $\tilde{X}_i$ and $\tilde{Y}_i$ respectively.

Then we just have to check

(a) The distribution for Alice and Bob is drawn from μ^1, and the distribution for Charlie and Dana is drawn from μ^n.

(b) Π' is correct if Π is correct: $\Pi'(U, V) = U \wedge V = \Pi(U, V)$

(c) $I((X_i, Y_i); \Pi | R_Z) = I((U, V); \Pi' | RW)$. For this wee need the $Z_2, ...Z_n$ to be independent of Z_1 (which is why the uniform distribution was useful).

This step was a central step because we took the larger problem and reduced it to an arguably simpler one.

7. Today's main task: we'll show that $CIC^1_\varepsilon(\Pi) = \Omega(1) > 0$. We know that this ought to be non-trivial since up to this point we've not used all the information we have about Π; we have essentially not yet used the fact that we're communicating.

3 Sketch

We want to ask: what is $CIC^1_\varepsilon(\Pi)$?

To think about this, we'll ignore the public randomness (for now), allow long transcripts and private randomness, and assume that Π solves $DISJ^1$ (i.e. AND) with error $\leq \varepsilon$.

Note that Π is really four different distributions corresponding to each possible combination of Alice and Bob's bits: so Π_{ab} is the distribution of the transcript if $U = a$ and $V = b$. Π_{ab} is a distribution of the sorts of messages Alice and Bob are sending each other. Also notice that on conditioning we only care about 3 of the 4 distributions, since our distribution shouldn't produce Π_{11}.

We can also notice that $I((U, V); \Pi_{uv} | W) = \frac{1}{2} I(V; \Pi_{0v}) + \frac{1}{2} I(U; \Pi_{u0})$.

We then have a problem: what prevents $\Pi_{00} =_d \Pi_{01} =_d \Pi_{10}$? Then the transcripts would not depend on the index and there would be no hope of doing anything. In the simplified case with no private randomness,

then the transcripts when both have 0s, or when one or both have 1s are indistinguishable. We know that since the last bit of the transcript is the output and the output of Π_{11} should be 1, it should be distinct from the other three. In other words, we must be able to accept Π_{11} with high probability and reject the other three with high probability.

We then have the following lemma:

Lemma 1. *For all transcripts τ,*

$$Pr[\Pi_{00} = \tau] \cdot Pr[\Pi_{11} = \tau] = Pr[\Pi_{01} = \tau] \cdot Pr[\Pi_{10} = \tau].$$

This lemma is straightforward to prove, so the proof has been omitted.

We want to be able to say that $\Pi_{00}, \Pi_{01}, \Pi_{10}$ must have something that distinguishes them, but this lemma does not quite seem to do this.

We will then use the Hellinger distance (defined rigorously below and not to be confused with entropy) to produce the following corollary:

Corollary 2 (Rectangle Property)**.** $H(\Pi_{00}, \Pi_{11}) = H(\Pi_{10}, \Pi_{01})$.

Exercise 3. *Use the lemma and the definition of Hellinger distance to prove the corollary.*

This statement means that under this measure of distance between distributions, one pair has the same distance as the other pair. Along with this notion of Hellinger distance, we will also use two other notions of distance, including the divergence. We can't go back and forth between all three of these notions of distance because two are bounded between 0 and 1, while the third (divergence) can be infinitely large.

We will want to use the idea that if we have an error, the distance between Π_{00}, Π_{11} is $\geq 1 - 2\varepsilon = \delta$ to imply that the Hellinger distance $H(\Pi_{00}, \Pi_{11})$ is large. Then also $H(\Pi_{01}, \Pi_{10})$ is large by the rectangle property. But then we'll use the fact that the Hellinger distance is a proper distance measure to use the triangle inequality to say that $H(\Pi_{01}, \Pi_{00}) + H(\Pi_{00} + \Pi_{10})$ is also large. This implies that we should be able to tell if the first or second bit is 0 or 1; i.e. we have some mutual information, and hopefully this mutual information ($I((U, V); \Pi_{uv}|W) = \frac{1}{2} I(V; \Pi_{0v}) + \frac{1}{2} I(U; \Pi_{u0})$) is large.

If we have public randomness, then we will have ε_R and δ_R, and our final mutal information will have some relationship with ε_R and δ_R.

4 Somewhat More Formally

In this section, we will provide a formal definition of Hellinger distance and discuss more formally what it meant for distances to be "large."

Definition 4 (Hellinger Distance)**.** *The Hellinger Distance $H(P, Q)$ for P, Q supported on Ω is*

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{\omega \in \Omega} (\sqrt{P(\omega)} - \sqrt{Q(\omega)})^2}$$

Notice that for all P, Q, $0 \leq H(P, Q) \leq 1$, $H(P, Q) = 0$ iff $P = Q$, and the triangle inequality holds: $H(P, Q) + H(Q, R) \geq H(P, R)$. We can also notice that this distance measure is related to the L^2 norm.

We can expand the Hellinger distance to get that

$$\begin{aligned} H(P, Q) &= \frac{1}{\sqrt{2}} \sqrt{\sum_{\omega \in \Omega} (\sqrt{P(\omega)} - \sqrt{Q(\omega)})^2} \\ &= \frac{1}{\sqrt{2}} \sqrt{\sum_{\omega \in \Omega} P(\omega) + \sum_{\omega \in \Omega} Q(\omega) - 2 \sum_{\omega \in \Omega} \sqrt{P(\omega) Q(\omega)}} \\ &= \sqrt{1 - \sum_{\omega \in \Omega} \sqrt{P(\omega) Q(\omega)}} \end{aligned}$$

This means that the distance is related to the inner product of the two distributions.

The rectangle property then follows from this definition and the lemma.

We also need the total variation distance (the most basic notion of distribution distance):

Definition 5 (Total Variation Distance)**.** *The total variation distance $\delta(P,Q)$ for P,Q supported on Ω is:* $\delta(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega}|P(\omega)-Q(\omega)| = \frac{1}{2}||P-Q||_1$

Notice that for all P,Q, $0 \le \delta(P,Q) \le 1$.

Exercise 6. *Show that the Hellinger distance and total variation distance are related as follows:*

$$H(P,Q)^2 \le \delta(P,Q) \le \sqrt{2}H(P,Q)$$

This statement means that if we have a moderate total variation distance, we have a moderate Hellinger distance. If we have a moderate Hellinger distance, we lose more switching to total variation distance.

We know that the total variation distance for the Π_{ab}s is $\delta(\Pi_{ab},\Pi_{11}) = \frac{1}{2}\sum_\tau|\Pi_{ab}(\tau)-\Pi_{11}(\tau)| \ge 1-2\varepsilon$ for all $(a,b)\in\{00,01,10\}$.

This implies that $H(\Pi_{00},\Pi_{11}) \ge \frac{\delta}{2}$, so by the rectangle property also $H(\Pi_{01},\Pi_{10}) \ge \frac{\delta}{2}$. From this, we can say that $\delta(\Pi_{01},\Pi_{10}) \ge \frac{\delta^2}{2}$. Then the triangle inequality says that $\delta(\Pi_{01},\Pi_{00}) + \delta(\Pi_{00},\Pi_{10}) \ge \frac{\delta^2}{2}$.

This raises the question: if two distributions are far from each other, can we say they are divergent?

Lemma 7 (Pinsker's Inequality)**.** *If $\delta(P,Q) \ge \delta$, then $D(P||Q) \ge 2\log_2 e(\delta(P,Q))^2$.*

We can get some intuition for Pinsker's inequality using an example.

Example 8. *Let $P = Bern(1/2)$ and $Q = Bern(1/2-\delta)$ then we know that $\delta(P,Q) = \delta$. We also know that $D(P||Q) = O(\delta^2)$.*

Now, we want to return to our quantity $I((U,V);\Pi_{uv}|W) = \frac{1}{2}I(V;\Pi_{0v}) + \frac{1}{2}I(U;\Pi_{u0})$.

We can recognize that

$$I(V,\Pi_{0v}) = D(\frac{1}{2}(\Pi_{00},0)+\frac{1}{2}(\Pi_{01},1)||\frac{1}{2}(\Pi_{00}+\Pi_{01})\times\text{Bern}(1/2))$$

and

$$I(U,\Pi_{u0}) = D(\frac{1}{2}(\Pi_{00},0)+\frac{1}{2}(\Pi_{10},1)||\frac{1}{2}(\Pi_{00}+\Pi_{10})\times\text{Bern}(1/2)).$$

Let's say that $P_0 = \frac{1}{2}(\Pi_{00},0)+\frac{1}{2}(\Pi_{01},1)$, $Q_0 = \frac{1}{2}(\Pi_{00}+\Pi_{01})\times\text{Bern}(1/2)$, $P_1 = \frac{1}{2}(\Pi_{00},0)+\frac{1}{2}(\Pi_{10},1)$, and $Q_1 = \frac{1}{2}(\Pi_{00}+\Pi_{10})\times\text{Bern}(1/2)$.

The fact that $\delta(\Pi_{01},\Pi_{00}) + \delta(\Pi_{00},\Pi_{10}) \ge \frac{\delta^2}{2}$ tells us that either $\delta(\Pi_{00},\Pi_{01}) \ge \frac{\delta^2}{4}$ or $\delta(\Pi_{00},\Pi_{10}) \ge \frac{\delta^2}{4}$.

This tells us that $\delta(P_0,Q_0) \ge \frac{\delta^2}{8}$ or $\delta(P_1,Q_1) \ge \frac{\delta^2}{8}$.

From this fact and Pinsker's inequality, we can say that $D(P_0||Q_0)$ or $D(P_1||Q_1) \ge \Omega(\frac{\delta^4}{64}) = \Omega(\delta^4)$. But this means that $I(V,\Pi_{0v})$ or $I(U,\Pi_{u0}) \ge \Omega(\delta^4)$, so $I((U,V);\Pi_{uv}|W) \ge \Omega(\delta^4)$.

Therfore, we have shown that $\delta(\Pi_{00},\Pi_{11}) \ge 1-2\varepsilon = \delta$ implies $I((U,V);\Pi_{uv}|W) \ge \Omega(\delta^4)$, which means that we can conclude that the disjointness problem requires $\Omega(n)$ communication.

Exercise 9. *Show that slightly more careful accounting of error terms allows us to reach this conclusion with public randomness.*