CS 229r Information Theory in Computer Science	March 08, 2016
Lecture 13	
Lecturer: Madhu Sudan	*Scribe: Wyatt Mackey*

1 Today's Class

- We will complete the proof of the direct sum theorem, $CC(f^{\otimes n}) \geq \sqrt{n}\Omega(CC(f) - 1)$
- Compression of low information protocols
  - Protocols, Trees, Priors
  - Correlated sampling
  - Path sampling
  - Analysis

2 Review

We showed last time that if $f^{\otimes n}$ has protocol with $\text{Inf} \leq I, \text{Com} \leq C$, then f has a protocol with $\text{Inf} \leq I/n, \text{Comm} \leq C$.

Today, we will show that if f has a protocol with $\text{Inf} = I$ and $\text{Comm} = C$, then it has a protocol with Comm growing $O(\sqrt{IC}\log C)$. We will do this by Compressing Interactive Communication.

3 Protocols as Trees with Priors

We can view protocols as trees, by beginning, for instance, with Alice speaking, and 0s or 1s take us down a particular path. Eventually we will reach a leaf. Call this protocol π. Our simulations, then, will do the following: Alice and Bob engage in some conversation following a protocol π'. At the end of this, Alice will output a long interaction π_A, and Bob π_B, and we want to guarantee that $\pi_A = \pi_B$ with high probability. Moreover, we want the distribution of $\pi_A, \pi_B \approx \pi$.

We have $(X, Y) \sim \mu$. Alice knows X, Bob knows $X \sim \mu_{X|Y}$. Given a node V owned by Bob, say that the probability of going right from this node is p_V. Then let P_V^A be the probability that Alice things we'll go right. Similarly, P_V^B is the probability Bob thinks he'll go right. Then if $\pi = \pi_1\pi_2...\pi_k$, where $k = C = \text{Comm.}$ of the protocol, then we know $p_V \leftrightarrow \pi_i|\pi_{<i}X, Y$, $p_V^A \leftrightarrow \pi_i|\pi_{<i}, X$, and $p_V^B \leftrightarrow \pi_i|\pi_{<i} < Y$.

Let's look at one half the information cost,

$$\begin{aligned}
\text{IC} &= I(\pi; Y|X) + I(\pi; X|Y) \\
&= \sum_{i=1}^{k} I(\pi; Y|X\pi_{<i}) + \sum_{i=1}^{k} I(\pi; X|Y\pi_{<i}) \\
&= \sum_{i=1}^{k} H_i(\pi_i|X\pi_{<i}) - H(\pi_i|XY\pi_{<i})
\end{aligned}$$

Set each of these summands to be v_i. Then we know $\sum v_i \leq I$, $I \ll C$, and each v_i is small since the average is I/C. Now consider the extreme case where $v_i = 0$ for all I. That means that the distributions on right and left are always identical, i.e. Alice and Bob always agree exactly on what coins should be tossed. This then just means we can use the shared randomness.

More precisely, if $\pi_i|X\pi_{<i} \stackrel{d}{=} \pi_i|XY\pi_{<i} \stackrel{d}{=} \pi_i|Y\pi_{<i}$ for all i then we use shared randomness to sample a leaf.

4 Correlated Sampling

Now, what if $I = \epsilon \neq 0$? That means Alice knows at the nodes what coins to toss, and Bob has a very good guess–say an off by ϵ error on his probability estimate. Thus when $TVD(\pi_I|X\pi_{<i}, \pi_i|XY\pi_{<i})$ is very small, we'll use "correlated sampling."

Suppose Alice wants to toss a coin X with $E[X] = P$, and Bob wants to toss a coin Y with $E[Y] = q$.

Claim 1 *We can get $Pr[X \neq Y] \leq |q - p|$ with zero communication.*

How would you design such a protocol? Take the real interval from 0 to 1. Using the common randomness, pick a random number $x \in [0, 1]$. If this is less than p, Alice calls is 0, otherwise 1, and similarly for Bob but with q. Then this is a protocol with shared randomness, giving exactly what was described.

We can extend this to random variables over any domain without much difficulty: Alice knows P and wants $X \sim P$ on Ω, and Bob knows Q and wants $Y \sim Q$ on Ω. Then we want $Pr[X \neq Y] = O(\text{TVD}(P, Q))$. Note that Alice does not know Q (or Bob P) and this makes the problem more interesting! The solution turns out to be simple and the idea is very similar to the above. Think of Ω as being a sequence of bits, one for each element, corresponding to "is this my element of Ω?" Then P, Q look like histograms on this space. The Total Variation Distance corresponds to the area that is under one curve, but not the other (clearly symmetric since the histograms are normalized to one, so it doesn't depend on our choice of which curve to be which).

The game we'll play, then, it to throw random dots on the histogram, between 0 and 1 on one axis, and corresponding to a point of Ω on the other. Alice will pick the first point that sits in her histogram, and Bob will do the same. Note we know whether they'll be equal as soon as a point falls below either of the curves. Then we have that the probability of a disagreement is

$$Pr[\text{Disagreement}] = \frac{2\,\text{TVD}(P, Q)}{1 + \text{TVD}(P, Q)}.$$

Returning to the Tree + Priors description, we're assuming Alice and Bob have a pretty good idea of what leaf they should be at. Why not just let each of them pick according to their distribution according to correlated sampling? Maybe this will work out. However, this relies on TVD being very small. However, we've only been told that IC is very small. How does this relate to TVD?

Let's look at the following extreme case: say that all the information goes from Alice to Bob. Bob's prior is going left with probability 1/2, and right with probability 1/2. Let Alice's prior be $X_i = 1$ go left with probability $\frac{1}{2} + \delta$, and for $X_i = 0$, go left with probability $\frac{1}{2} - \delta$. Then the information cost is

$$= nD\left(\frac{1}{2}||\frac{1}{2} - \delta\right) = n\delta^2 = \epsilon.$$

On the other hand, what is the chance that Alice and Bob would agree on a leaf at the end? This corresponds to them ever disagreeing. We know

$$Pr[\text{Disagreement}] \simeq n\delta$$

if $\delta << \frac{1}{n}$. However, this is larger that $n\delta^2$ by a factor of $\frac{1}{\delta}$, so it turns out that

$$Pr[\text{Disagreement}] \simeq \sqrt{n\epsilon} = \sqrt{CI}.$$

This is the right number that we were looking for, but in entirely the wrong place. This, however, is the correct intuition to use for the remainder of our discussion.

Finally, we must conisder the case where the information cost is very large. For each node, Alice will toss a coin biased to p_V^A, and Bob will do similarly with bias p_V^B. They will toss coins in the correllated manner described above. Then each person thinks there is some path which we're following. At some node they diverged, and the person who made the choice at the node is the person who made the "right choice" in some sense. Then we want to figure out where Alice and Bob diverged, which we'll do using path sampling.

5 Path Sampling

Continue in the game described just before in the previous section. Starting at the current root, for every node V below, sample the outoing edge A with probability p_V^A, B with protability p_V^B, $Pr[\text{Disagreement}] \leq |p_V^A p_V B|$. This yields leaves ℓ_A, ℓ_B. Let $v =$ least common ancestor of ℓ_A, ℓ_B. If $v =$ this leaf, we're done. Otherwise, repeat with v being the new root, taking the correct choice here to go one step below.[1] The claim is that at this point, the output is correct/follows the correct distribution, but we still need to show that it has low information cost.

First question: how do we find the least common ancestor with little communication? Can we even tell if we're in the same place with little communication? Yes, we know that we can do equality communication with constant complexity. Then we can do a binary search to find the first point at which this equality no longer holds. Again, equality is constant cost, so determining v will take $O(\log C)$ communication. If we put a factor of log on the theorem, we can get this dealing with all kinds of extra events in $O(\log^3 C)$ communication instead, and this will be sufficient.

Now on this "right path," we can ask how many errors there are. Let Z_i be the random variable that is 1 if there is a disagreement at level i, and 0 otherwise. Then the number of iterations will be $\sum Z_i$, and we therefore want to bound $E[\sum Z_i] = \sum E[Z_i]$. We know this is just

$$\sum \text{TVD}(\pi_i|X\pi_{<i}, \pi_i|Y\pi_{<i}).$$

On the other hand, we already have a bound on the information complexity

$$I = \sum_{i=1}^{k} |H(\pi_i|X\pi_{<i}) - H(\pi_i|Y\pi_{<i})|.$$

Note technically we should have both X and Y in the second entropy, not just Y. However, if it's Bob's choice, they're the same, and if it's Alice, then this would be the first term and the absolute

[1] We don't actually have to go down one step here, and the presentation in class did not include this originally. While it's cleaner to do it this way, it is unnecessary and was not the exposition chosen in class.

value corrects it. Again, call each of these summands v_i. Then Pinsker's inequality tells us the TVD will be bounded by this divergence, so

$$\begin{aligned}\sum \mathrm{TVD}(\pi_i|X\pi_{<i}, \pi_i|Y\pi_{<i}) &\leq \sum O(\sqrt{v_i}) \\ &\leq \sqrt{k}\, O\left(\sqrt{\sum v_i}\right)\end{aligned}$$

by Cauchy-Schwarz, which we know is $\sqrt{C}O(\sqrt{I})$.

It is perhaps possible to tighten this down by being a bit more careful in our work, such as maybe $O(\sqrt{IC})$. There has also been some work to not have this go to infinity with C: eventually, Braverman showed that for every protocol with Inf I, there is a protocol with Comm being 2^I. Ganor, Kol, and Raz recently showed there is a problem with inf. cost $\leq I$ and comm. cost $\geq 2^I$.

6 Aside

Consider the following game/puzzle/challenge, due to Dana Moshkovitz: We have a bag full of coins, zeroes and ones on the two sides, with the promise that $\frac{2}{3}$ of the coins are $\geq .9$ biased, so they come up with 1 at least .9 of the times. We have a parameter n, and the challenge is to find a $\geq .7$ biased coin with probability of error being exponential in $-n$, with $O(n)$ coin tosses. (It is known that this problem can be solved with $O(n \log n)$ coin tosses.)