CS 229r Information Theory in Computer Science March 14, 2019

Lecture 14

Instructor: Madhu Sudan *Scribe: Alec Sun*

Today the instructor will just ramble on about random topics until the time expires. This is because he does not want to start a new topic before spring break. They will be tangentially related to the breadth of projects that are available to choose, and maybe today's lecture will help you choose your final project topic.

We are currently in the middle of information complexity, and this itself is a central theme of the course.

Definition 1. *Recall that the information complexity of a protocol is*

$$\mathrm{IC}_{ext}(\Pi) = I(xy; \Pi).$$

This is known in the literature is the external information complexity*. We then call the* internal information complexity *as*

$$\mathrm{IC}_{int}(\Pi) = I(y; \Pi \mid x) + I(x; \Pi \mid y)$$

Alice knows x so we don't want to count the information that she already knew, likewise for Bob.

Proposition 2. *We have*

$$\mathrm{IC}_{int}(\Pi) \leq \mathrm{IC}_{ext}(\Pi) \leq \mathrm{CC}(\Pi).$$

Why should we define the internal communication complexity? The answer is that it is more representative of what is limiting communication between multiple parties. This leads to one of the important papers we will cover after spring break due to Barak et. al., in which they show the following theorem:

Theorem 3. *If Π communicates $C = \mathrm{CC}(\Pi)$ bits and reveals $I = \mathrm{IC}_{int}(\Pi)$ bits of information, then the protocol can be compressed to a $\tilde{O}(\sqrt{IC})$-bit protocol.*

Remark What this theorem implies is that in such a situation, every unit of time players are sending each other vacuous bits and only rarely are sending a useful bit of information. But yet the theorem does not squish communication down to the entropic limit. This theorem tells us that I is indeed related to communication complexity.

Remark The reason why we cannot keep applying the theorem again and again to reduce the number of bits used in the protocol is that in the new protocol with less bits, the internal information leaked is now more, so I increases.

Later, a subset of the authors of this paper, Braverman and Rao, gives a different theorem. It tells us a statement about amortized communication complexity.

Theorem 4 (Amortized Communication Complexity)**.** *Suppose we draw samples $x^{(1)}, x^{(2)}, \ldots$ and $y^{(1)}, y^{(2)}, \ldots$ and consider the protocol Π on these inputs*

$$\Pi(x^{(1)}, y^{(1)}), \ldots, \Pi(x^{(t)}, y^{(t)}).$$

It turns out that amount of communication needed to convey t samples is bounded by

$$tI \leq \textit{communication needed} \leq tC.$$

But in fact this theorem tells us that we can get the bound

$$tI \leq \textit{communication needed} \leq O(tI).$$

This theorem basically tells us that amortized communication complexity approximated information complexity. Hence it is possible to get a result in information complexity that dominates both of these complexities. Then came a tour-de-force work by Ganor, Kol, and Raz that tells us that we cannot produce a theorem that implies both types of information complexities.

Theorem 5. *There exists a function f such that the information cost satisfies $\mathrm{IC}_{int}(f) = k$ but the communication complexity satisfies $\mathrm{CC}(f) \geq 2^k$.*

Hence we can not get the squishing that we might hope for and can only get an exponential gap. This is contrasted by the following theorem by Braverman:

Theorem 6. *For all functions f, we have*

$$\mathrm{CC}(f) < 2^{\mathrm{IC}(f)}.$$

In the early days of information theory it used to be the case that identical results were published in the United States and Russia. Now we don't have that problem, but we have a different problem. We have a lot of computer scientists working on such problems, but the subject primarily resides in a community of information theorists. In particular, the amortized communication complexity theorem above seems to also be the same as Distributed Source Coding by Ishwar and Ma. The interaction between the two literatures is very interesting. Computer science tends to own the communication complexity viewpoint and information theorists tend to own the information theoretic aspect.

Recall that

$$\mathrm{IC}(f) = \min_{\Pi \text{ computes } f} \mathrm{IC}(\Pi).$$

By focusing on mutual information we get the t samples out of the system, but we have a new problem: how many bits should Π be communicating? How would you determine this minimum? The space of all protocols is countably infinite. Maybe communication complexity on a single instance is very hard to figure out. We can say the same thing about entropy. If we want to compress a whole ensemble of random variables, it is fortunate that both single-shot and amortized entropy are polynomial time computable. For single-shot we can use Huffman Coding. If we go beyond entropy, however, we run into challenges. For example, if we look at channel capacity, a single-shot compression could in theory be a **NP**-complete problem. That being said, for amortized this happens to be a convex space so we can apply convex optimization and it turns out to have a polynomial time algorithm! In other words, many samples could potentially be easier than single-shot.

We have discussed protocols in which we are allowed make an error $\varepsilon \to 0$. But what if we required there to be absolutely no error? This turns out to be a different beast. For zero-error channel capacity, we know that it is **NP**-complete. But we do not know whether or not it is

computable, and we do not know whether of not it is in **P**either. This is known as the *Shannon capacity* of a graph. One of the lovely papers in this field is due to Lovasz, who came up with a novel way to study this problem. Zero-error is a very combinatorial type of question, and $\varepsilon \to 0$ error is an information theoretic question.

We now talk about another information theoretic problem.

Definition 7. *In the* common randomness generation *problem, Alice gets x and Bob gets y as an input where (x, y) is jointly distributed from a distribution μ with possible correlation. Suppose $x = x_1, \ldots, x_t$ and $y = y_1, \ldots, y_t$. The goal is for Alice and Bob to output $r_1, \ldots, r_{\rho t}$ bits that are hopefully uniformly random. Let the number of bits communicated be γt.*

Does the underlying distribution μ permit protocols with (γ, ρ)? Certainly if Alice gets private randomness r_A and Bob get r_B they can transmit their private randomness to each other to create random bits, but we would like protocols for which $\gamma \ll \rho$.

Remark What are the random bits produced by the protocol random to? The answer is the observer. If they are totally random to the observer, then this is known as *secret key generation.* If we have less communication than the number of random bits being generated, then this implies, with converse, that we can get a secret key out of the protocol. We subtract the amount of information in the protocol itself from the rest.

Remark When $x = y$ or when x, y are independent this problem turns out to be easy to analyze. It is when x, y are slightly correlated that this becomes a hard problem.

There is a remarkable paper by Witsenhausen. What if we first perform some operations that transforms initial randomness into something else?

Example 8. *Suppose we have the distribution with associated probabilities*

$$(x, y) = \begin{cases} 00 & \textit{probability } \frac{1}{2} \\ 01 & \textit{probability } \frac{1}{4} \\ 10 & \textit{probability } \frac{1}{4} \end{cases}.$$

Suppose Alice outputs r_1 and Bob outputs r_1'. Then the pair (r_1, r_1') consists of usually equal bits but occasionally not.

Following this paper came another work by Ahlswede and Csiszar. This paper answers the following question: what is

$$\max_{\Pi} \frac{\mathrm{IC}_{\mathrm{ext}}(\Pi)}{\mathrm{IC}_{\mathrm{int}}(\Pi)}?$$

Internal information makes sense because it is the amortized communication complexity. But this ratio also turns out to be very fundamental in the line of research. Some of the project topics relate to this ratio between external and internal information complexities.

Remark The information complexity $\mathrm{IC}(f)$ is not known to be computable but is known to be *computably approximable.* That is, an ε-approximation can be computed in polynomial time. There are theorems by Braverman that tell us which joint distributions μ of (x, y) are computably approximable, equivalently for which (γ, ρ) is the information complexity computably approximable?

Another question surrounding feasibility of computability is computing the entropy of a Markov chain. In the two-state example we gave with a noisy state and a stable state, we still do not know yet of an polynomial time computable approximation algorithm for the entropy of this particular source, let alone Hidden Markov Models in general!

Lastly we will try to touch on some topics that relate to information theory but can often constitute entire classes or departments on their own.

- Streaming algorithms
  - Communication lower bounds lead to streaming lower bounds. The algorithms are limited in both speed and memory in this setting. When we put algorithms under such severe restrictions their ability to perform computation is limited. How can we capture the lower bound limits?
- Data structures
  - We have a massive amount of information. We want to preprocess it and store it in a reasonable amount of space so that we can query certain types of questions from it efficiently.
  - The limits here are the amount of space and the query time. This combination of time and space is very similar to the two limits that appear in streaming algorithm.
  - The information theory lower bounds also lead to data structure lower bounds.
- Differential privacy
  - How is privacy defined in a probabilistic sense? There are some underlying distributions that we want to retrieve information from.
  - There are a fair amount of questions here where information theoretic tools are used to design mechanisms but also to analyze limits.
  - There are some suggestions for project over here but the topic is very vast.
- Learning, Statistics, and Finance
  - Sad.
- Optimization
  - Information theory has the ability to help us with the complexity of optimization.
  - There is an area called *extension complexity* started by Yannakakis. It was a response to a prank in the community regarding someone who claimed to prove $\mathbf{P} = \mathbf{NP}$. The researcher kept saying: "Oh! This is actually correct but you forgot this one technicality." He kept coming up with more and more example of how the traveling salesman problem can be transformed to linear programming. Yannakakis realized that all you are trying to do is consider some high dimensional object like in the traveling salesman problem such that if you project it down to n-dimensional space you get a polytope. Then you are analyzing Hamiltonian walks on this polytope. All of these papers essentially reduced to this polytope, but Yannakakis proved that the high dimensional object needs exponentially many constraints to create the polytope. Hence this disproved all of these $\mathbf{P} = \mathbf{NP}$ pranks.

  - The paper by Yannakakis uses lower bounds from information complexity, and goes from talking about optimization to geometry and finally communication complexity.
  - This field deals with restricted impossibility results. The general question remains: "Is there a complicated non-symmetric linear program that can express things like the traveling salesman problem?" This was finally resolved recently by Fiorini et. al., and they use Set Disjointness lower bounds to give an answer in the negative.
  - Continuing in the way, there has been more work by Braverman and Moitra that uses information complexity to establish more lower bounds.
  - This collection of works is wonderful to trace, and somewhere in the history there should be an accessible, beautiful result.

- Hardness of approximation
  - How do you know whether or not there is not an algorithm that approximates a problem very well?
  - The idea for answering this question is something called *2-prover proof systems* as well as the idea of *parallel repetition*
  - This is highly related to one of the paradigms in information theory: you define some version of a question related to a single instance of a problem, and then consider what happens when you try to solve many instances at the same time.
  - This field was started by Raz in 1994. We will definitely talk about this subject in lecture. Ever since this paper, the proofs have been improved significantly by Holenstein and others. Some of the projects points us toward these papers. But the instructor says that out of all of the applications, hardness of approximation is probably the most difficult to wrap your head around.

- Lower bounds in property testing

- Topics related to things we have already discussed
  - Hardness of Gap Hamming problem, related to hardness of Set Disjointness
  - Channel coding
  - Things related to Shearer's Lemma. Recall that this says that

    $$\text{3D volume} \leq \sqrt{\text{product of 2D areas}}.$$

    A paper by Ellis et. al. discusses perturbations in the 3D object and discusses what might be equal to

    $$\sqrt{\text{product of 2D areas}}.$$

  - Constructive proof of Lovasz Local Lemma using entropy