CS 229r Information Theory in Computer Science Jan 26, 2016

Lecture 1

Lecturer: Madhu Sudan *Scribe: Badih Ghazi*

1 Course Overview

The goal of this course is to learn some basic information theory, and to illustrate its mathematical power by studying its applications in combinatorics, complexity theory, algorithms, privacy, etc.. The course will be run in seminar-style. In the first few lectures, Madhu will cover some basic information theory background, and the rest of the lectures will be student-run. The grades will be based on projects, presentations, participation and scribe work.

2 Motivating Example: Shearer's Lemma

On a high-level, Shearer's Lemma can be thought of as a combinatorial analog of the Loomis-Whitney inequality, a geometric inequality relating the volume of a 3D object to areas of its 2D projections. In Shearer's Lemma, we will be given sets in a finite universe, and we will be talking about cardinalities instead of volumes and areas.
We start with some notation. Let $[n] := \{1, 2, 3, \ldots, n\}$. For a subset $S = \{i_1, i_2, \ldots, i_k\} \subseteq [d]$ and $x = (x_1, x_2, \ldots, x_d) \in [n]^d$, we let $x_S := (x_{i_1}, x_{i_2}, \ldots, x_{i_k})$. For a collection $F \subseteq [n]^d$ and a subset $S \subseteq [d]$, denote

$$F_S := \{x_S \mid x \in F\}.$$

We now give the statement of Shearer's lemma.

Theorem 1 (Shearer's Lemma; (k,d)-version) *Let $F \subseteq [n]^d$, then*

$$|F|^{\binom{d-1}{k-1}} \leq \prod_{S \subseteq [d] : |S| = k} |F_S|.$$

In this rest of this section, we sketch the proof of Theorem 1. We start with the particular case where $d = 1$. Note that since

$$F \subseteq F_{\{1\}} \times F_{\{2\}} \times F_{\{3\}} \times \ldots \times F_{\{k\}},$$

we have that

$$|F| \leq |F_{\{1\}}| \cdot |F_{\{2\}}| \cdot |F_{\{3\}}| \cdot \ldots \cdot |F_{\{k\}}|.$$

The next simplest special case that we can consider is the one where $d = 2$ and $k = 3$. In this case, the $(3,2)$-version of Theorem 1 claims that

$$|F|^2 \leq |F_{\{1,2\}}| \cdot |F_{\{2,3\}}| \cdot |F_{\{3,1\}}|.$$

Proving this particular case of Theorem 1 is already non-trivial!

We next show how to use *information theory* in order to prove Theorem 1. We point out that the presented proof (as well as today's lecture in general) will be handwavy. The formal definitions will be introduced in the future lectures.

We first consider the following setup:

- You and I know F.
- I am given $W \in F$.
- I need to describe W to you.
- How many bits do I need to send to you ?

Roughly, we will define the *entropy* $H(W)$ of W to measure this quantity. (Note that the entropy depends not only on F, but also on how W is chosen from F).

Suppose further that $W = (X, Y, Z)$ and you and I already know X, but Eve doesn't know Z. How many bits does Eve expect me to send you ? We will define this quantity to be the *conditional entropy* of W given Z.

We now consider the following "axioms of (conditional) entropy":

- If $w \in F$, then $H(W) \leq \log_2(|F|)$.
- If w is drawn uniformly at random from F, and if $|F| = 2^j$, then $H(W) = j = \log_2(|F|)$.
- $H(X|Y) \leq H(X)$. (Intuitively, this says that in the worst-case, you and I can ignore Y and focus on sending X. So X given Y is easier to communicate than X alone).
- $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

It is a remarkable fact (which we will formally see in the next lecture) that there does exist an operator $H(X)$ satisfying the above axioms.

Going back to the proof of Shearer's lemma (Theorem 1 above), let's first revisit the particular case where $d = 1$ using the language of information theory. Let (X, Y) be chosen uniformly at random from F. Then,

- $H(X, Y) = \log_2(|F|)$.
- $\log_2(|F_{\{1\}}|) \geq H(X)$ (note that we don't have equality since projecting (X, Y) on X doesn't not necessarily give a uniform distribution on $F_{\{1\}}$).
- Similarly, $\log_2(|F_{\{2\}}|) \geq H(Y)$.

Thus, it suffices to show that $H(X, Y) \leq H(X) + H(Y)$. Note that this follows from the "axioms" that $H(X, Y) = H(X) + H(Y|X)$ and $H(Y|X) \leq H(Y)$. We now illustrate the proof of the general (k, d)-version of Shearer's lemma by working out the proof in the particular case where $k = 3$ and $d = 2$. We wish to prove that

$$|F|^2 \leq |F_{\{1,2\}}| \cdot |F_{\{2,3\}}| \cdot |F_{\{3,1\}}|.$$

This would follow if we can establish

$$2H(X, Y, Z) \leq H(X, Y) + H(Y, Z) + H(X, Z). \tag{1}$$

Note that we have that

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

On the other hand, we have that

$$2H(X,Y,Z) = 2H(X) + 2H(Y|X) + 2H(Z|X,Y).$$

Using the "axioms" that $H(Y|X) \leq H(Y)$, $H(Z|X,Y) \leq H(Z|X)$ and $H(Z|X,Y) \leq H(Z|Y)$, we conclude Equation (1), and hence the $(3,2)$-version of Shearer's Lemma. The more general (k,d)-version is left as an exercise.

The moral of the story is that in the above axioms lies a very interesting inequality, which can be very useful. Information theory has many other interesting notions (mutual information, divergence, Hellinger distance) and inequalities (Fano, Pinsker). In this course, we will study these notions, inequalities and their applications. Notable applications of information theory in computer science include parallel repetition, communication complexity, data structures, streaming algorithms, optimization, cryptography, privacy, etc..

3 Administrative Stuff

In the next two lectures, Madhu will review some information theory basics. Afterwards, students will jointly present the lectures with Madhu. Each lecture will have a paper and a student covering it. The student will read the paper one week before, and privately present it to Madhu. The audience will skim over the paper and come in. The first paper will be "Entropy and Counting" by Jaikumar Radhakrishnan (a link to this paper is available on the course website).