CS 229r Information Theory in Computer Science Jan 29, 2019

Lecture 1

Instructor: Madhu Sudan *Scribe: Kenz Kallal*

1 Welcome to CS221

1.1 Course Information

Contact information and office hours:

- **Lecturer:** Madhu Sudan (madhu@cs.harvard.edu).
- **Prof. Sudan's Office Hours:** Tuesday, Thursday 1:15–2:15.
- **TF:** Mitali Bafna (mitali.bafna@gmail.com)
- **TF's Office Hours:** This week Wednesday and Friday 4:30–5:30 at LISE 319.

1.2 Course Expectations

Grades will be based on the following:

- 3–5 problem sets
- Scribing ≥ 1 lecture
- Final project
- Participation (in class and on Piazza)

1.3 Course Topics

The first few lectures will be about the basics of information theory. Then, they will cover applications of information theory to computer science:

- Limits on the performance of data structures
- How well can information be compressed?
- Error-correcting codes
- Communication complexity
- Streaming
- Differential privacy
- Optimization

2 Basics of Information Theory

Today we will not be rigorous about the definitions or manipulations of notions from information theory. Instead, we will give a sense of how the tools of information theory might be applied to solve interesting problems.

2.1 Random Variables

Let X be a random variable with probability distribution P_X. In this context it is convenient to restrict X to a compact set Ω. Recall that random variables X, Y can be jointly distributed with probability distribution P_{XY}. This carries the data $\{P_{Y|X=\alpha}\}_{\alpha\in\Omega}$ of probability distributions for Y given any possible fixed value of X.

2.2 Entropy

Today we will not give a fully rigorous definition of entropy, but the following "definition" will suffice to motivate our use of it in the next section.

Definition 1 (Entropy)**.** Let X be a random variable. The *entropy* of X, denoted $H(X)$, is "the number of bits needed, in expectation, to convey X."

For example, Alice and Bob might both know P_X, and they need to come up with a protocol to compress X and send it over the line to each other.

So far we have no rigorous way to calculate the entropy of a random variable, but intuition tells us what the answers are in some easy examples:

Example 2. *Suppose P_X is the uniform distribution over $\{0,1\}^n$. Then intuitively we must use n bits to convey X, and we can write "$H(X) = n$"*

Example 3. *Suppose X is 0^n with probability $1/2$ and is uniformly distributed over $\{0,1\}^n$ with probability $1/2$. Then we can use a single bit to indicate which case occurs, and an additional n bits in case the second case occurs. The expected value of the number of bits used is*

$$\frac{1+(n+1)}{2} = \frac{n}{2} + 1$$

so we can write "$H(X) \approx n/2$."

Definition 4 (Conditional Entropy)**.** The entropy of Y conditioned on X, denoted $H(Y|X)$, is "the number of bits needed, in expectation, to convey Y given that X is known." More precisely,

$$H(Y|X) = \mathop{\mathbb{E}}_{\alpha\in\Omega}[H(Y|X=\alpha)]$$

where $Y|X=\alpha$ is distributed according to the joint distribution P_{XY}.

Example 5. *Suppose X and Y are independent and uniformly distributed over $\{0,1\}^n$. Intuitively, knowing X does not give any additional information about Y, we can write "$H(Y|X) = H(Y) = n$."*

Example 6. *Suppose X is uniformly distributed over $\{0,1\}^n$, and Y is uniformly distributed over $\{0,1\}^{2n}$ such that X consists of the first n bits of Y. Then, given X, one knows the first n bits of Y (and no other information is conveyed by X) so we can write "$H(Y|X) = n$."*

We now state some intuitive axioms for entropy.

(1) If $|\Omega| < \infty$, then $H(X) \le \log|\Omega|$ with equality if and only if P_X is uniform on Ω.

(2) $H(X,Y) = H(X) + H(Y|X)$. "to specify X and Y it suffices to specify X and then Y given that X has already been transmitted." One can show that this method of transmitting X, Y is optimal. [NB: this axiom is frequently called the *chain rule* of conditional entropy]

(3) $H(Y|X) \leq H(Y)$.

Warning: axiom (3) does not necessarily work when specialized to an arbitrary value of X; it is only true in expectation over all possible values of X.

Exercise 7. *Construct a counterexample to the assertion that $H(Y|X = \alpha) \leq H(Y)$.*

3 Shearer's Lemma

Let $F \subseteq [N]^d$ represent some object in d-dimensional space. For any set $S \subseteq [d]$ with $|S| = k \leq d$, we can project F to a k-dimensional object on the coordinates described by S. In particular, if $S = \{i_1, \ldots, i_k\}$ with WLOG $i_1 < \cdots < i_k$, we can define

$$F_S := \{(x_{i_1}, \ldots, x_{i_k}) : (x_1, \ldots, x_d) \in F\}.$$

Intuitively, knowing that the projections of F are small should tell us that F cannot be too big. This is the content of Shearer's Lemma.

Lemma 8 (Shearer's Lemma)**.** *Let $F \subseteq [N]^d$ and $k \leq d$. Then*

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

In the case $d = 3, k = 1$, this specializes to the following:

Lemma 9 (Shearer's Lemma, "infant version")**.** *Let $F \subseteq [N]^3$. Then*

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

Proof. Each element of F is of the form (x_1, x_2, x_3), where by definition $x_i \in F_{\{i\}}$ for $i = 1, 2, 3$. So, we have an inclusion of sets

$$F \subseteq F_{\{1\}} \times F_{\{2\}} \times F_{\{3\}}.$$

Taking the cardinalities of both sides, the result is immediate. □

We use entropy to prove a harder case, namely $d = 3, k = 2$.

Lemma 10 (Shearer's Lemma, "baby version")**.** *Let $F \subseteq [N]^3$. Then*

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

Proof. Take the random variable (X, Y, Z) to be uniformly distributed on F. By Axiom (1), we know

$$H(X, Y, Z) = \log |F|.$$

By definition of the projections,

- (X, Y) is restricted to $F_{\{1,2\}}$
- (Y, Z) is restricted to $F_{\{2,3\}}$
- (X, Z) is restricted to $F_{\{1,3\}}$

So Axiom (1) yields

$$\begin{aligned} H(X,Y) &\le \log|F_{\{1,2\}}| \\ H(Y,Z) &\le \log|F_{\{2,3\}}| \\ H(X,Y) &\le \log|F_{\{1,3\}}|. \end{aligned}$$

To show the desired result $|F|^2 \le |F_{\{1,2\}}||F_{\{2,3\}}||F_{\{1,3\}}|$, by taking logs it therefore suffices to show

$$2H(X,Y,Z) \le H(X,Y) + H(Y,Z) + H(X,Y).$$

Using Axiom (2), we have

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

Axiom (3) tells us that $H(Y) \ge H(Y|X)$ and $H(Z|X), H(Z|Y) \ge H(Z|X,Y)$. Adding up the three equations above and applying these inequalities,

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

The right hand side is equal to $2H(X,Y) + 2H(Z|X,Y) = 2H(X,Y,Z)$ by axiom (2), which yields the desired result.

$\square$

Exercise 11. *Can the proof of the baby version of Shearer's Lemma be extended to the general case?*