CAS CS 591 B: Communication Complexity

Prof. Mark Bun — Fall 2019

Lecture Notes 7:

Introduction to Information Complexity

Reading.

- Rao-Yehudayoff Chapter 6

1 A Few More Facts about Mutual Information

Lemma 1.

$$I(A;B) = I(A;B|C) - I(A;C|B) + I(A;C)$$

Proof. By applying the chain rule two different ways,

$$I(A;BC) = I(A;C) + I(A;B|C) = I(A;B) + I(A;C|B).$$

Rearranging gives the identity. □

The following claim states that post-processing cannot increase the amount of information one random variable reveals about another. (Think, e.g., of $C = f(B)$ for a randomized function f, where the randomness in f is independent of everything else.)

Lemma 2 (Information Processing Inequality)**.** *Let $A \to B \to C$ be a Markov chain, i.e., C is independent from A conditioned on B. Then $I(A;C) \leq I(A;B)$.*

Proof.

$$I(A;C) \leq I(A;C) + I(A;B|C) = I(A;BC) = I(A;B) + I(A;C|B) = I(A;B),$$

where the last equality follows because $I(A;C|B) = 0$ by definition. □

2 KL Divergence

Mutual information gives us a way of measuring how far two random variables are from being independent. A related way to measure similarity between two distributions is via the notion of KL divergence, or relative entropy.

Let A, B be random variables over the *same* sample space X, with probability mass functions p and q, respectively. Then we define

Definition 3 (KL Divergence)**.**

$$D(A\|B) = \sum_{x \in X : p(x) > 0} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim A} \log \frac{p(x)}{q(x)}$$

Like entropy, KL divergence has an intuitive interpretation in terms of coding messages. We may write

$$D(A\|B) = \mathbb{E}_{x\sim A} \log \frac{1}{q(x)} - \mathbb{E}_{x\sim A} \log \frac{1}{p(x)}.$$

The second term is simply the entropy of A, i.e., the minimum expected length for an encoding of A. The first term we can think of as the expected length of an encoding of A using a code which was optimized for B. So KL divergence captures the loss of using a code designed for B rather than a code designed for A.

Let's record a few facts about KL divergence:

- $D(A\|A) = 0$
- $D(A\|B) \geq 0$. This follows by Jensen:

$$\begin{aligned} -\mathbb{E}_{x\sim A} \log \frac{p(x)}{q(x)} &\leq -\log \mathbb{E}_{x\sim A} \frac{p(x)}{q(x)} \\ &= \log \mathbb{E}_{x\sim A} \frac{q(x)}{p(x)} \\ &= \log 1 = 0. \end{aligned}$$

- Unlike mutual information, $D(A\|B) \neq D(B\|A)$ in general
- If A is supported on a point outside of the support of B, then $D(A\|B) = \infty$

KL divergence also has a nice connection to mutual information. For jointly distributed random variables A, B, let $A \otimes B$ denote the product distribution with marginals A and B.

Theorem 4. $I(A;B) = D(AB\|A \otimes B)$

Proof.

$$\begin{aligned} D(AB\|A \otimes B) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x)} + \frac{1}{p(y)} - \frac{1}{p(x,y)} \right) \\ &= H(A) + H(B) - H(AB) = I(A;B). \end{aligned}$$

□

3 Information Cost

We are now ready to use information theory to define notions of information cost for communication protocols. Let μ be a distribution over $X \times Y$ and let Π be a communication protocol using private randomness R_A and R_B for Alice and Bob respectively, and public randomness R. Define the transcript of the protocol, denoted $T(x, y, r_A, r_B, r)$, to consist of the public randomness string followed by the sequence of messages exchanged between Alice and Bob. Let AB be jointly distributed over $X \times Y$ according to μ.

The first way one might model the information cost of a protocol is in terms of how much information about Alice and Bob's inputs is revealed to an external observer.

Definition 5. The external information cost of a protocol Π with respect to a distribution μ, denoted $IC_\mu^{\text{ext}}(\Pi) = I(AB;T)$.

It turns out that there is another way to model information cost which is easier to work with. This is the notion of *internal information cost*, or the amount Alice and Bob learn about each others' inputs over the course of the protocol.

Definition 6. The (internal) information cost of a protocol Π with respect to a distribution μ, denoted $IC_\mu(\Pi) = I(A;T|B) + I(B;T|A)$.

You may (very reasonably) ask why the conditioning does not include Alice and Bob's private randomness. This is because conditioning on private randomness does not affect the information cost:

Lemma 7. $I(A;T|BR_B) = I(A;T|B)$

Proof. We prove the claim by induction on the number of rounds of the protocol. Let $T_{\leq k}$ denote the prefix of the transcript through round k. The claim is clearly true for T_0, which consists only of the public randomness. Suppose it is Bob's turn to speak in round k and the claim is true through round $k-2$, so our induction hypothesis says

$$I(A;T_{\leq k-1}|BR_B) = I(A;T_{\leq k-1}|B).$$

By applying the chain rule twice,

$$\begin{aligned} I(A;T_{\leq k}|B) &= I(A;T_{\leq k-1}|B) + I(A;T_k|BT_{\leq k-1}) \\ &= I(A;T_{\leq k-2}|B) + I(A;T_{k-1}|BT_{\leq k-2}) + I(A;T_k|BT_{\leq k-1}) \end{aligned}$$

Similarly, we can write

$$I(A;T_{\leq k}|BR_B) = I(A;T_{\leq k-2}|BR_B) + I(A;T_{k-1}|BR_BT_{\leq k-2}) + I(A;T_k|BR_BT_{\leq k-1}).$$

The third term of each identity is zero, since it's Bob's turn to speak in round k: Alice's input does not affect T_k except through $T_{\leq k-1}$ which is already being conditioned on. By the induction hypothesis, it's enough to show that

$$I(A;T_{k-1}|BT_{\leq k-2}) = I(A;T_{k-1}|BR_BT_{\leq k-2}).$$

To see this, we use Lemma 1 to write

$$I(A;T_{k-1}|BT_{\leq k-2}) = I(A;T_{k-1}|BT_{\leq k-2}R_B) - I(R_B;T_{k-1}|BT_{\leq k-2}A) + I(R_B;T_{k-1}|BT_{\leq k-2}).$$

The last two terms are zero because it is Alice's turn to speak in round $k-1$; hence, Bob's randomness R_B does not affect T_{k-1} except through $T_{\leq k-2}$. □

4 Information vs. Communication

We establish the following relationships showing that information lower bounds communication. Let $CC_\mu(\Pi)$ denote the expected number of bits exchanged under Π for inputs chosen from μ.

Theorem 8. *For every distribution μ,*

$$IC_\mu(\Pi) \leq IC_\mu^{\text{ext}}(\Pi) \leq CC_\mu(\Pi).$$

Proof. We begin with the first inequality. By repeatedly applying the chain rule, we get

$$IC_\mu^{\text{ext}}(\Pi) = I(AB;T) = \sum_{k=1}^{L} I(AB;T_k|T_{\leq k-1}),$$

where L is the length of the protocol. By the chain rule and non-negativity of mutual information, we have

$$I(AB;T_k|T_{\leq k-1}) \geq \max\left\{I(A;T_k|BT_{\leq k-1}), I(B;T_k|AT_{\leq k-1})\right\}.$$

Now observe that if it is Alice's turn to speak in round k, then $I(B;T_k|AT_{\leq k-1}) = 0$. Similarly, if it's Bob's turn to speak, then $I(A;T_k|BT_{\leq k-1}) = 0$. So we actually have

$$I(AB;T_k|T_{\leq k-1}) \geq I(A;T_k|BT_{\leq k-1}) + I(B;T_k|AT_{\leq k-1}).$$

Repeatedly applying the chain rule again lets us conclude

$$\begin{aligned} IC_\mu^{\text{ext}}(\Pi) &= \sum_{k=1}^{L} I(AB;T_k|T_{\leq k-1}) \\ &\geq \sum_{k=1}^{L} I(A;T_k|BT_{\leq k-1}) + I(B;T_k|AT_{\leq k-1}) \\ &= I(A;T|B) + I(B;T|A) \\ &= IC_\mu(\Pi). \end{aligned}$$

Next, to bound the external information by the communication, we simply observe:

$$IC_\mu^{\text{ext}}(\Pi) = I(\Pi_0;AB) + I(\Pi_{>0};AB|\Pi_0) = I(\Pi_{>0};AB|\Pi_0) \leq H(\Pi_{>0}),$$

and the latter is at most the average total length of messages sent between Alice and Bob.

□