

INFORMATION THEORY CONTINUED

(Last Lecture)

Reference: • Communication and Information Complexity
Mark Braverman

Recap:

① Entropy: $H(X) = \mathbb{E}[I(X)] = \sum_x p(x) \cdot I(x)$, where $I(x) = \log_2\left(\frac{1}{p(x)}\right)$

② Some properties of entropy:

a) For $|\text{range}(X)| \leq N$, $0 \leq H(X) \leq \log_2 N$, with
 $H(X) = 0$ iff X is constant
 $H(X) = \log_2 N$ iff X is uniform r.v.

b) For independent r.v.s X & Y $H(XY) = H(X) + H(Y)$

③ Mutual Information: $I(X; Y) = H(X) + H(Y) - H(XY)$

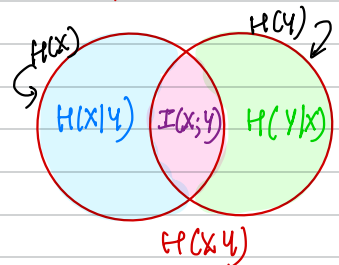
④ Properties of mutual information:

a) $I(X; Y) = 0$ iff X & Y are independent.

b) Conditional Entropy

$$H(Y|X) = H(Y) - I(X; Y) \quad (\geq 0)$$

$$H(X|Y) = H(X) - I(X; Y) \quad (\geq 0)$$



c) $I(X; X) = H(X)$

$$I(X; Y) = I(Y; X)$$

d) Chain Rule: $H(XY) = H(X) + H(Y|X)$

Generalized Chain rule: $H(X_1 X_2 \dots X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1 \dots X_{n-1})$

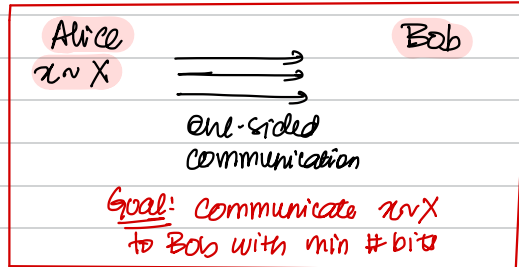
e) $H(Y|X) = \mathbb{E}_{x \sim X} [H(Y|X=x)]$

⑤ Conditional mutual information: $I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(XY|Z)$
 $= \mathbb{E}_{z \sim Z} [I(X|Z=z; Y|Z=z)]$

INFORMATION COMPLEXITY

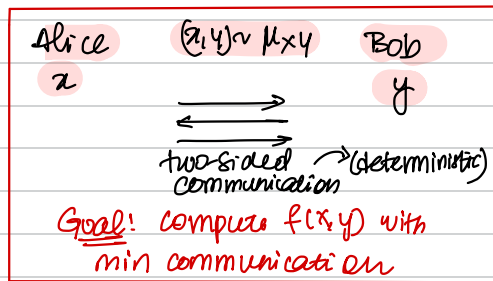
(Application to comm-complexity)

View Shannon's entropy as follows:



This one-sided communication complexity $\geq H(X)$

We want to define an analogue to $H(X)$ for the two sided communication:



Recall:

$D_{\mu_{X,Y}}^{\epsilon}(f)$: min communication cost over all deterministic protocols π s.t.
 $\text{err}_{\mu}(w) = \Pr_{(x,y) \sim \mu_{X,Y}} [\pi(x,y) \neq f(x,y)] \leq \epsilon$

Definition (Information Complexity)

Information complexity of a function f captures the least amount of information that Alice & Bob need to exchange about their inputs to successfully compute f on their i/p's.

- ① Information cost of a protocol π under $\mu_{x,y}$
 let T_π denote the random variable corresponding to the transcript of protocol π
 (all messages communicated in $\pi(x,y)$)

$$IC_\mu(\pi) := \underbrace{I(Y; T_\pi | X)}_{\text{Info Alice learns about } Y \text{ from the protocol}} + \underbrace{I(X; T_\pi | Y)}_{\text{Info Bob learns about } X \text{ from the protocol.}}$$

Remark: the distribution μ impacts $IC_\mu(\pi)$. To give you a simple example: If μ is singleton distribution supported on one input (x_0, y_0) , then $IC_\mu(\pi) = 0 \forall \pi$. (why?)
Example connecting to Shannon entropy: Recall in one-way communication $X \sim \mu$ & $Y = \perp$.

$$IC_\mu(\pi) = I(X; T_\pi) = H(X) - \underbrace{H(X | T_\pi)}_{=0} = H(X).$$

↓
Only Bob trying to get X from T_π

- ② Information complexity of f under $\mu_{x,y}$

$$IC_\mu^b(f) := \infimum \{ IC_\mu(\pi) : \pi \text{ computes } f \text{ with } err_\mu(\pi) \leq \epsilon \}$$

Why not minimum?

This is a subtle point (try to recall the difference between min & inf)
 if you understand then read below.

communication protocols of a given comm. cost is finite, but same is not true about information cost (think why?)
 i.e. one can have sequence $\pi_1, \pi_2, \dots$ of protocols of ever-increasing comm. cost, but whose information complexity $IC_\mu(\pi_n)$ converges to $IC_\mu^b(f)$ in limit.

ADDITIVITY OF INFORMATION COMPLEXITY

[Bar-Yossef, Jayram, Kumar, Sivakumar '04]

Let $f^{\otimes n}$ denote the function:

$$f^{\otimes n}((x_1, \dots, x_n), (y_1, \dots, y_n)) = (f(x_1, y_1), \dots, f(x_n, y_n))$$

$\mu^{\otimes n}$ denotes the product distribution, i.e.

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \sim \mu \text{ (independently)}$$

Theorem I: $IC_{\varepsilon}^{\mu^{\otimes n}}(f^{\otimes n}) = n \cdot IC_{\varepsilon}^{\mu}(f)$.

Pf (sketch) will explain for $n=2$ (you can generalize easily)

Claim 1: $IC_{\varepsilon}^{\mu \otimes \mu}(f \otimes f) \leq 2 \cdot IC_{\varepsilon}^{\mu}(f)$

Proof: Let π be any arbitrary protocol computing f .
Run two independent copies of π to compute $f \otimes f$.

Let (x_1, y_1) & (x_2, y_2) denote the independent input distributions for the two independent runs, denoted by π_1 & π_2 . Apply chain rule & prove (exercise)

$$\textcircled{1} \quad I(y_1, y_2; \underbrace{T_{\pi_1 \pi_2}}_{\text{transcript of 2 rounds}} | x_1, x_2) = I(y_1; T_{\pi_1} | x_1) + I(y_2; T_{\pi_2} | x_2)$$

$$\& \textcircled{2} \quad I(x_1, x_2; T_{\pi_1 \pi_2} | y_1, y_2) = I(x_1; T_{\pi_1} | y_1) + I(x_2; T_{\pi_2} | y_2)$$

$$\textcircled{1} \& \textcircled{2} \Rightarrow IC_{\mu \otimes \mu}(\pi_1, \pi_2) = IC_{\mu_1}(\pi_1) + IC_{\mu_2}(\pi_2)$$

why? Justify

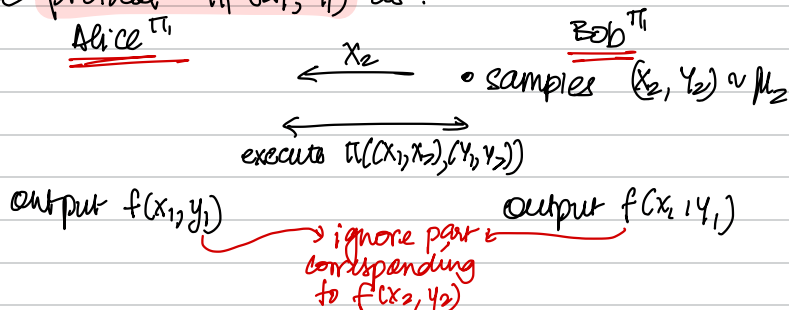
$$\Rightarrow IC_{\varepsilon}^{\mu \otimes \mu}(f \otimes f) \leq IC_{\varepsilon}^{\mu}(f) + IC_{\varepsilon}^{\mu}(f) = 2 \cdot IC_{\varepsilon}^{\mu}(f)$$

Claim 2: $IC_{\varepsilon}^{\mu \otimes \mu}(f \otimes f) \geq 2 \cdot IC_{\varepsilon}^{\mu}(f)$

Proof: Now, we are given a protocol π computing $f \otimes f$ & need to construct two protocols π_1 & π_2 (computing f) such that $IC_{\mu_1}(\pi_1) + IC_{\mu_2}(\pi_2) \leq IC_{\mu \otimes \mu}(\pi)$, where

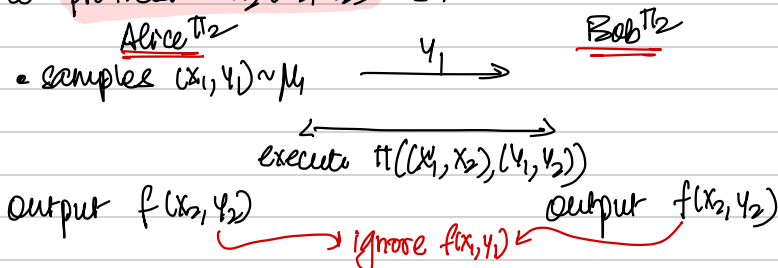
μ_1 & μ_2 are two independent copies of μ .

Define protocol $\pi_1(x_1, y_1)$ as:



$$\text{Now } IC_{\mu_1}(\pi_1) = I(Y_1; T_{\pi} | X_1 X_2) + I(X_1; T_{\pi} | X_2 Y_1 Y_2) \quad \rightarrow \textcircled{1}$$

Define protocol $\pi_2(x_2, y_2)$ as:



Similar to above

$$IC_{\mu_2}(\pi_2) = I(Y_2; T_{\pi} | Y_1 X_1 X_2) + I(X_2; T_{\pi} | Y_1 Y_2) \quad \rightarrow \textcircled{2}$$

$\textcircled{1}, \textcircled{2}$ & chain rule gives

$$\begin{aligned} & IC_{\mu_1}(\pi_1) + IC_{\mu_2}(\pi_2) \\ &= I(X_1; T_{\pi} | X_2 Y_1 Y_2) + I(X_2; T_{\pi} | Y_1 Y_2) + I(Y_1; T_{\pi} | X_1 X_2) + I(Y_2; T_{\pi} | Y_1 X_1 X_2) \\ &= I(X_1 X_2; T_{\pi} | Y_1 Y_2) + I(Y_1 Y_2; T_{\pi} | X_1 X_2) \\ &= IC_{\mu \otimes \mu}(\pi) \end{aligned}$$

Again justify this

$$\Rightarrow \underbrace{IC_{\mu_1}^n(f) + IC_{\mu_2}^n(f)}_{2 \cdot IC_{\mu}^n(f)} \geq IC_{\mu \otimes \mu}^n(f \otimes f)$$

claims $\textcircled{1}$ & $\textcircled{2}$ complete the proof for $n=2$
 $\rightarrow$ generalize for n

It is easy (and intuitive) to see that

$$D_{\varepsilon}^M(f) \geq IC_{\varepsilon}^M(f)$$

($\therefore$ the information cost of a protocol π will always be bounded by size of $cost(\pi)$)

$$\therefore \frac{1}{n} \cdot D_{\varepsilon}^{M^{\otimes n}}(f^{\otimes n}) \geq \frac{1}{n} \cdot IC_{\varepsilon}^{M^{\otimes n}}(f^{\otimes n})$$

$$= \frac{1}{n} \cdot n \cdot IC_{\varepsilon}^M(f) \quad (\text{By Theorem 1})$$

$$\text{i.e. } \frac{1}{n} \cdot D_{\varepsilon}^{M^{\otimes n}}(f^{\otimes n}) \geq IC_{\varepsilon}^M(f)$$

DISJOINTNESS_n(x, y)

$$DISJOINTNESS_n(x, y) := \bigwedge_{i=1}^n AND_2(x_i, y_i)$$

Recall from lecture 20: $R_{\varepsilon, \text{pub}}(DISJOINTNESS_n) = \max_{\mu} D_{\varepsilon}^M(DISJOINTNESS_n)$

$\therefore$ To lower bound $R_{\varepsilon, \text{pub}}(DISJOINTNESS_n)$, we need to find some "hard" distribution μ & lower-bound $D_{\varepsilon}^M(DISJOINTNESS_n)$.

Intuitively, $DISJOINTNESS_n(x, y)$ will have similar communication complexity as $AND_2^{\otimes n}(x, y)$. Hence, we will bound $D_{\varepsilon}^{M^{\otimes n}}(AND_2^{\otimes n})$.

$\therefore$ We want a "hard" $\mu^{\otimes n}$.

Recall, we saw that if $\mu^{\otimes n}$ was uniform distribution, then it is an easy distribution for Alice & Bob (see lecture 20). In fact, if many (x_i, y_i) 's are $= (1, 1)$ (which happens when $\mu(1, 1)$ is high) then w.h.p. Alice & Bob know the output is 0 making it an easy distribution.

Try a few more attempts for μ on your own.

Conclusion: A hard distribution μ is such that:

$$\mu(1, 1) \sim \frac{1}{n} \rightarrow 0$$

We won't see a proof of the following theorem (It's an amazing result $\rightarrow$ read the paper if interested)

★ no proof in this course.

Theorem [Braverman, Garg, Pankratov, Weinstein '13]

$$D_{\epsilon}^{\mu, \text{com}}(\text{DISJOINTNESS}_n) \geq n \cdot \left(\underbrace{IC_0^{\mu}(\text{AND}_2)}_{\rightarrow \text{as } \epsilon \rightarrow 0} - o(1) \right)$$

for any μ with $\mu((1,1)) = 0$.

i.e. want to now understand 0-error information complexity of the two-bit AND-function.

It is non-trivial to actually show that $IC_0^{\mu}(\text{AND}_2) \approx 0.4\epsilon$

Corollary $\forall \epsilon > 0, \mu^*$ such that

$$D_{\epsilon}^{\mu^*}(\text{DISJOINTNESS}_n) \geq 0.4\epsilon n - o(n)$$

$\rightarrow 0 \text{ as } \epsilon \rightarrow 0$

$\Rightarrow$ By Yao's Principle, we get

$$R_{\epsilon, \text{pub}}(\text{DISJOINTNESS}_n) \geq 0.4\epsilon n - o(n)$$

You can refer to Braverman's notes and the original paper (if interested) to understand the complete proof.

Another amazing result (capturing our intuition) is (with stats without proof)

Amortized communication cost = Information complexity

[Braverman-Rao '11]

$$\lim_{n \rightarrow \infty} \frac{1}{n} D_{\epsilon}^{\mu^{\otimes n}}(f^{\otimes n}) = IC_{\epsilon}^{\mu}(f)$$

(Hence, $IC_{\epsilon}^{\mu}(f)$ truly is analogue of $H(x)$ in the two-sided communication setting.)

That's all Folks!! Thank You