Information Theory for Communication Complexity

David P. Woodruff
IBM Almaden

Talk Outline

1. Information Theory Concepts

2. Distances Between Distributions

3. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

4. Communication Lower Bounds imply space lower bounds for data stream algorithms

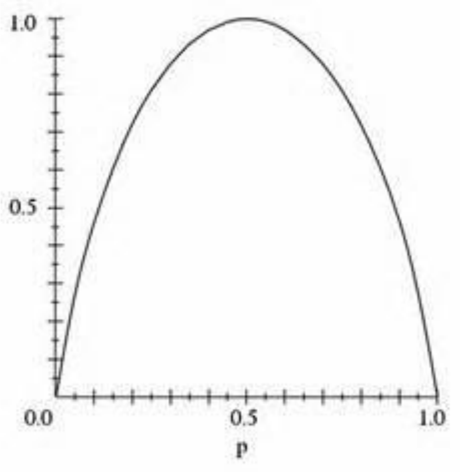
5. Techniques for Multi-Player Communication

Discrete Distributions

- Consider distributions p over a finite support of size n:
  - p = $(p_1, p_2, p_3, \dots, p_n)$
  - p_i ∈ [0,1] for all i
  - $\sum_i p_i$ = 1
- X is a random variable with distribution p if $\Pr[X = i] = p_i$

Entropy

- Let X be a random variable with distribution p on n items
- (Entropy) $H(X) = \sum_i p_i \log_2 (1/p_i)$
  - If $p_i = 0$ then $p_i \log_2 \left(\frac{1}{p_i}\right) = 0$
  - $H(X) \leq \log_2 n$. Equality holds when $p_i = \frac{1}{n}$ for all i.
  - Entropy measures "uncertainty" of X.



- (Binary Input) If B is a bit with bias p, then

$$H(B) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

(symmetric)

Conditional and Joint Entropy

- Let X and Y be random variables

- (Conditional Entropy)

  $$H(X \mid Y) = \sum_y \mathrm{H(X \mid Y = y)\ Pr}[Y = y]$$

- (Joint Entropy)

  $$\mathrm{H(X, Y)} = \sum_{x,y} P\mathrm{r}[(\mathrm{X,Y}) = (\mathrm{x,y})] \log(1/\mathrm{Pr}[(\mathrm{X,Y}) = (\mathrm{x,y})])$$

Chain Rule for Entropy

- (Chain Rule) H(X,Y) = H(X) + H(Y | X)

- Proof:

H(X,Y) = $\sum_{x,y} \Pr[(X,Y) = (x,y)] \log\left(\frac{1}{\Pr((X,Y)=(x,y))}\right)$

= $\sum_{x,y} \Pr[X = x\,] \Pr\,[Y = y\,|X = x] \log\left(\frac{1}{\Pr(X=x)\,\Pr(Y=y\,|X=x)}\right)$

= $\sum_{x,y} \Pr[X = x\,] \Pr\,[Y = y\,|X = x](\log\left(\frac{1}{\Pr(X=x)}\right) + \log(\frac{1}{\Pr[Y=y\,|X=x]}))$

= H(X) + H(Y | X)

Conditioning Cannot Increase Entropy

- Let X, Y be random variables. Then H(X | Y) $\leq H(X)$

- To prove this, we need Jensen's Inequality:

continuous

Recall a concave function f means $f\left(\frac{a+b}{2}\right) \geq \frac{f(a)}{2} + \frac{f(b)}{2}$ for all a,b

Recall the expectation $E[W] = \sum_w \Pr[W = w] \cdot w$

(Jensen's Inequality) For concave f, $E[f(W)] \leq f(E[W])$

We will use that $f(x) = \log(x)$ is concave

Conditioning Cannot Increase Entropy

- Proof:

$$H(X\,|Y) - H(X) = \sum_{xy} \Pr[Y = y] \Pr[X = x \,|\, Y = y] \log(\frac{1}{\Pr[X=x \,|Y=y]})$$

$$- \sum_x \Pr[X = x] \log\left(\frac{1}{\Pr[X=x]}\right) \sum_y \Pr[Y = y \,|\, X = x]$$

$$= \sum_{x,y} \Pr[X = x, Y = y] \log(\frac{\Pr[X=x]}{\Pr[X=x \,|Y=y]})$$

$$= \sum_{x,y} \Pr[X = x, Y = y] \log(\frac{\Pr[X=x]\Pr[Y=y]}{\Pr[(X,Y)=(x,y)]})$$

$$\leq \log(\sum_{x,y} \Pr[X = x, Y = y] \cdot \frac{\Pr[X=x]\Pr[Y=y]}{\Pr[(X,Y)=(x,y)]})$$

$$= 0$$

where the inequality follows by Jensen’s inequality.

If X and Y are independent H(X | Y) = H(X).

Mutual Information

- (Mutual Information) $I(X ; Y) = H(X) - H(X \mid Y)$

$$= H(Y) - H(Y \mid X)$$

$$= I(Y ; X)$$

Note: $I(X ; X) = H(X) - H(X \mid X) = H(X)$

- (Conditional Mutual Information)

$$I(X ; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

Chain Rule for Mutual Information

- I(X, Y ; Z) = I(X ; Z) + I(Y ; Z | X)

- Proof: I(X, Y ; Z) = H(X, Y) – H(X, Y | Z)

  = H(X) + H(Y | X) – H(X | Z) – H(Y | X, Z)

  = I(X ; Z) + I(Y; Z | X)

By induction, $I(X_1, \ldots, X_n; Z) = \sum_i I(X_i; Z \mid X_1, \ldots, X_{\{i-1\}})$

Fano’s Inequality

- For any estimator X’: X -> Y -> X’ with $P_e = \Pr[X' \neq X]$, we have

$$H(X\,|Y) \leq H(P_e) + P_e \cdot \log(|X| - 1)$$

Here X -> Y -> X’ is a Markov Chain, meaning X’ and X are independent given Y.

“Past and future are conditionally independent given the present”

To prove Fano’s Inequality, we need the data processing inequality

Data Processing Inequality

- Suppose X -> Y -> Z is a Markov Chain. Then,
$$I(X\,;Y) \geq I(X;Z)$$
- That is, no clever combination of the data can improve estimation

- I(X ; Y, Z) = I(X ; Z) + I(X ; Y | Z) = I(X ; Y) + I(X ; Z | Y)
- So, it suffices to show I(X ; Z | Y) = 0
- I(X ; Z | Y) = H(X | Y) – H(X | Y, Z)
- But given Y, then X and Z are independent, so H(X | Y, Z) = H(X | Y).

- Data Processing Inequality implies H(X | Y) $\leq H(X\,|Z)$

Proof of Fano's Inequality

- For any estimator X' such that X-> Y -> X' with $P_e = \Pr[X \neq X']$,

we have $H(X\,|Y) \leq H(P_e) + P_e(\log_2|X| - 1)$.

Proof: Let E = 1 if X' is not equal to X, and E = 0 otherwise.

H(E, X | X') = H(X | X') + H(E | X, X') = H(X | X')

H(E, X | X') = H(E | X') + H(X | E, X') ≤ $H(P_e)$ + H(X | E, X')

But H(X | E, X') = Pr(E = 0)H(X | X', E = 0) + Pr(E = 1)H(X | X', E = 1)

$$\leq (1 - P_e) \cdot 0 + P_e \cdot \log_2(|X| - 1)$$

Combining the above, H(X | X') ≤ $H(P_e) + P_e \cdot \log_2(|X| - 1)$

By Data Processing, H(X | Y) ≤ $H(X\,|X') \leq H(P_e) + P_e \cdot \log_2(|X| - 1)$

Tightness of Fano's Inequality

- Suppose the distribution p of X satisfies $p_1 \geq p_2 \geq \ldots \geq p_n$
- Suppose Y is a constant, so I(X ; Y) = H(X) – H(X | Y) = 0.
- Best predictor X' of X is X = 1.
- $P_e = \Pr[X' \neq X] = 1 - p_1$
- H(X | Y) $\leq H(p_1) + (1 - p_1)\log_2(\mathrm{n} - 1)$ predicted by Fano's inequality
- But H(X) = H(X | Y) and if $p_2 = p_3 = \ldots = p_n = \frac{1-p_1}{n-1}$ the inequality is tight

Talk Outline

1. Information Theory Concepts

2. Distances Between Distributions

3. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

4. Communication Lower Bounds imply space lower bounds for data stream algorithms

5. Techniques for Multi-Player Communication

Distances Between Distributions

- Let p and q be two distributions with the same support
- (Total Variation Distance) $D_{TV}(p,q) = \frac{1}{2}|p-q|_1 = \frac{1}{2}\sum_i |p_i - q_i|$
  - $D_{TV}(p,q) = \max_{events\ E} |p(E) - q(E)|$
- Sometimes abuse notation and say $D_{TV}(X,Y)$ to mean $D_{TV}(p,q)$ where X has distribution p and Y has distribution q
- (Hellinger Distance)
  - Define $\sqrt{p} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n})$, $\sqrt{q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_n})$
  - Note that $\sqrt{p}$ and $\sqrt{q}$ are unit vectors
  - h(p,q) = $\frac{1}{\sqrt{2}}|\sqrt{p}$ - $\sqrt{q}|_2 = \frac{1}{\sqrt{2}}\left(\sum_i(\sqrt{p_i} - \sqrt{q_i})^2\right)^{.5}$
- Note: $D_{TV}(p,q)$ and $h(p,q)$ satisfy the triangle inequality

Why Hellinger Distance?

- Useful for independent distributions

- Suppose X and Y are independent random variables with distributions p and q, respectively

$$\Pr[(X,Y)=(x,y)]=p(x)\cdot q(y)$$

- Suppose A and B are independent random variables with distributions p' and q', respectively

$$\Pr[(A,B)=(a,b)]=p'(a)\cdot q'(b)$$

- (Product Property)

$$\mathrm{h}^2\big((X,Y),(A,B)\big)=1-(1-h^2(X,A))\cdot(1-h^2(Y,B))$$

No easy product structure for variation distance

Product Property of Hellinger Distance

- $h^2\big((p,q),(p',q')\big) = \frac{1}{2}\left|\sqrt{p,q} - \sqrt{p'q'}\right|_2^2$

$$= \frac{1}{2}\left(1 + 1 - 2\left\langle\sqrt{p,q}, \sqrt{p'q'}\right\rangle\right)$$

$$= 1 - \sum_{i,j} \sqrt{p_i}\sqrt{q_j}\sqrt{p_i'}\sqrt{q_j'}$$

$$= 1 - \sum_i \sqrt{p_i}\sqrt{p_i'} \cdot \sum_j \sqrt{q_j}\sqrt{q_j'}$$

$$= 1 - (1 - h^2(p,p')) \cdot (1 - h^2(q,q'))$$

Jensen-Shannon Distance

- (Kullback-Leibler Divergence) KL(p,q) = $\sum_i p_i \log\left(\frac{p_i}{q_i}\right)$
  - KL(p,q) can be infinite!

- (Jensen-Shannon Distance) JS(p,q) = $\frac{1}{2}(KL(p,r) + KL(q,r))$, where r = (p+q)/2 is the average distribution

- Why Jensen-Shannon Distance?

- (Jensen-Shannon Lower Bounds Information) Suppose X, B are possibly dependent random variables and B is a uniform bit. Then,
$$I(X;B) \geq JS(X \mid B=0, X \mid B=1)$$

Relations Between Distance Measures

- (Squared Hellinger Lower Bounds Jensen-Shannon)

$$JS(p,q) \geq h^2(\mathrm{p,q})$$

- (Squared Hellinger Lower Bounded by Squared Variation Distance)

$$h^2(\mathrm{p,q}) \geq D_{TV}^2(p,q)$$

- (Variation Distance Upper Bounds Distinguishing Probability)

If you can distinguish distribution p from q with a sample w.pr. δ,

$$D_{TV}(p,q) \geq \delta$$

Talk Outline

1. Information Theory Concepts

2. Distances Between Distributions

3. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

4. Communication Lower Bounds imply space lower bounds for data stream algorithms

5. Techniques for Multi-Player Communication

Randomized 1-Way Communication Complexity

INDEX
PROBLEM

x 2 $\{0,1\}^n$

j 2 $\{1, 2, 3, \ldots, n\}$

- Alice sends a single message M to Bob
- Bob, given M and j, should output x_j with probability at least 2/3
- Note: The probability is over the coin tosses, not inputs
- Prove that for some inputs and coin tosses, M must be $\Omega(n)$ bits long…

1-Way Communication Complexity of Index

- Consider a uniform distribution μ on X
- Alice sends a single message M to Bob
- We can think of Bob's output as a guess $X_j' to\ X_j$
- For all j, $\Pr\left[X_j' = X_j\right] \geq \frac{2}{3}$

- By Fano's inequality, for all j,

$$H\left(X_j \middle| M\right) \leq H\left(\frac{2}{3}\right) + \frac{1}{3}(\log_2 2\ - 1) = H(\frac{1}{3})$$

1-Way Communication of Index Continued

- Consider the mutual information I(M ; X)
- By the chain rule,

$$I(X ; M) = \Sigma_i\, I(X_i ; M \mid X_{<i})$$
$$= \Sigma_i\, H(X_i \mid X_{<i}) - H(X_i \mid M , X_{<i})$$

- Since the coordinates of X are independent bits, $H(X_i \mid X_{<i}) = H(X_i) = 1$.
- Since conditioning cannot increase entropy,

$$H(X_i \mid M , X_{<i}) \leq H(X_i \mid M)$$

So, $I(X ; M) \geq n - \sum_i H(X_i|M) \geq n - H\left(\frac{1}{3}\right) n$

So, $|M| \geq H(M) \geq I(X ; M) = \Omega(n)$

Talk Outline

1. Information Theory Concepts

2. Distances Between Distributions

3. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

4. Communication Lower Bounds imply space lower bounds for data stream algorithms

5. Techniques for Multi-Player Communication