Outline

1. Information Theory Concepts

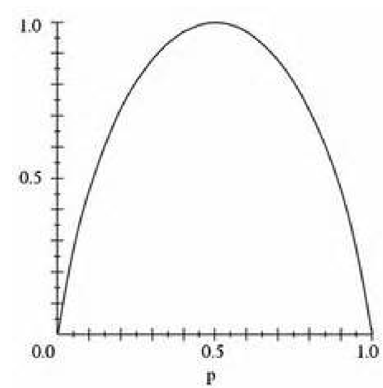
2. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

Discrete Distributions

- Consider distributions p over a finite support of size n:
  - p = $(p_1, p_2, p_3, \ldots, p_n)$
  - $p_i \in [0,1]$ 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)

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

- (Joint Entropy)

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

Chain Rule for Entropy

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

- Proof:

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

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

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

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

Conditioning Cannot Increase Entropy

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

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

Let f be a continuous, concave function, and let $p_1, \ldots, p_n$ be non-negative reals that sum to 1. For any $x_1, \ldots, x_n$,

$$\sum_{i=1,\ldots,n} p_i f(x_i) \leq f(\sum_{i=1,\ldots,n} p_i x_i)$$

- Recall that f is concave if $f\left(\frac{a+b}{2}\right) \geq \frac{f(a)}{2} + \frac{f(b)}{2}$ and f(x) = log x is concave

Conditioning Cannot Increase Entropy

- Proof:

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

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

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

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

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

$$= 0$$

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

Is $I(X ; Y \mid Z) \geq I(X ; Y)$? Or is $I(X ; Y \mid Z) \leq I(X; Y)$? Neither!

Mutual Information

- Claim: For certain X, Y, Z, we can have I(X ; Y | Z) ≤ I(X ; Y)

- Consider X = Y = Z

- Then,
  - $\mathrm{I(X\,;Y\mid Z) = H(X\mid Z)\ - H(X\mid Y,Z) = 0\ - 0 = 0}$
  - $\mathrm{I(X\,;Y) = H(X) - H(X\mid Y) = H(X) - 0 = H(X)}$

- Intuitively, Y only reveals information that Z has already revealed, and we are conditioning on Z

Mutual Information

- Claim: For certain X, Y, Z, we can have I(X ; Y | Z) $\geq$ I(X ; Y)
- Consider $\mathrm{X} = \mathrm{Y} + \mathrm{Z} \bmod 2$, where X and Y are uniform in {0,1}
- Then,
  - $\mathrm{I}(\mathrm{X}\,;\mathrm{Y} \mid \mathrm{Z}) = \mathrm{H}(\mathrm{X} \mid \mathrm{Z}) \ - \mathrm{H}(\mathrm{X} \mid \mathrm{Y}, \mathrm{Z}) = 1 \ - 0 = 1$
  - $\mathrm{I}(\mathrm{X}\,;\mathrm{Y}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X}\,|\mathrm{Y}) = 1 \ - 1 = 0$
- Intuitively, Y only reveals useful information about X after also conditioning on 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) $\leq H(X \mid Z)$

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

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

Tightness of Fano's Inequality

- Suppose the distribution p of X satisfies $p_1 \geq p_2 \geq \ \dots \geq p_n$
- 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 $p_2 = p_3 = \ \dots = p_n = \frac{1-p_1}{n-1}$ the inequality is tight

Tightness of Fano's Inequality

- For X from distribution $(p_1, \frac{1-p_1}{n-1}, \dots, \frac{1-p_1}{n-1})$
- $H(X) = \sum_i p_i \log\left(\frac{1}{p_i}\right)$

$$= p_1 \log\left(\frac{1}{p_1}\right) + \sum_{i>1} \frac{1-p_1}{n-1} \log(\frac{n-1}{1-p_1})$$

$$= p_1 \log\left(\frac{1}{p_1}\right) + (1-p_1) \log\left(\frac{1}{1-p_1}\right) + (1-p_1)\log(n-1)$$

$$= H(p_1) + (1-p_1)\log(n-1)$$

Talk Outline

1. Information Theory Concepts

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

Randomized 1-Way Communication Complexity

INDEX
PROBLEM

$x \in \{0,1\}^n$ $j \in \{1, 2, 3, \ldots, n\}$

- 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,

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

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

Typical Communication Reduction

$a \in \{0,1\}^n$
Create stream s(a)

$b \in \{0,1\}^n$
Create stream s(b)

Lower Bound Technique

1. Run Streaming Alg on s(a), transmit state of Alg(s(a)) to Bob

2. Bob computes Alg(s(a), s(b))

3. If Bob solves g(a,b), space complexity of Alg at least the 1-way communication complexity of g

Example: Distinct Elements

- Given $a_1, \ldots, a_m$ in $[n]$, how many *distinct* numbers are there?

- Index problem:
  - Alice has a bit string x in $\{0, 1\}^n$
  - Bob has an index i in $[n]$
  - Bob wants to know if $x_i = 1$

- Reduction:
  - $s(a) = i_1, \ldots, i_r$, where i_j appears if and only if $x_{i_j} = 1$
  - $s(b) = i$
  - If $Alg(s(a), s(b)) = Alg(s(a))+1$ then $x_i = 0$, otherwise $x_i = 1$

- Space complexity of Alg at least the 1-way communication complexity of Index

Strengthening Index: Augmented Indexing

- Augmented-Index problem:
  - Alice has $x \in \{0, 1\}^n$
  - Bob has $i \in [n]$, and $x_1, \ldots, x_{i-1}$
  - Bob wants to learn x_i

- Similar proof shows $\Omega(n)$ bound
- $I(M ; X) = \text{sum}_i \, I(M ; X_i \mid X_{< i})$

  $= n - \text{sum}_i \, H(X_i \mid M, X_{< i})$

- By Fano's inequality, $H(X_i \mid M, X_{< i}) \leq H(\delta)$ if Bob can predict X_i with probability $\geq 1- \delta$ from $M, X_{< i}$
- $CC_\delta(\text{Augmented-Index}) \geq I(M ; X) \geq n(1-H(\delta))$

Log n Bit Lower Bound for Estimating Norms

- Alice has $x \in \{0,1\}^{\log n}$ as an input to Augmented Index
- She creates a vector v with a single coordinate equal to $\sum_j 10^j x_j$
- Alice sends to Bob the state of the data stream algorithm after feeding in the input v
- Bob has i in [log n] and $x_{i+1}, x_{i+2}, \ldots, x_{\log n}$
- Bob creates vector w = $\sum_{j>i} 10^j x_j$
- Bob feeds –w into the state of the algorithm
- If the output of the streaming algorithm is at least $10^i/2$, guess $x_i = 1$, otherwise guess $x_i = 0$

$\frac{1}{\epsilon^2}$ Bit Lower Bound for Estimating Norms

$x \in \{0,1\}^n$ $y \in \{0,1\}^n$

- Gap Hamming Problem: Hamming distance $\Delta(x,y) > n/2 + 2\varepsilon n$ or $\Delta(x,y) < n/2 + \varepsilon n$
- Lower bound of $\Omega(\varepsilon^{-2})$ for randomized 1-way communication [Indyk, W], [W], [Jayram, Kumar, Sivakumar]
- Gives $\Omega(\varepsilon^{-2})$ bit lower bound for approximating any norm
- Same for 2-way communication [Chakrabarti, Regev]

Gap-Hamming From Index [JKS]

Public coin = $r^1, \ldots, r^t$, each in $\{0,1\}^t$

$t = \Theta(\epsilon^{-2})$

$x \in \{0,1\}^t$ → $a \in \{0,1\}^t$

$i \in [t]$ → $b \in \{0,1\}^t$

$a_k = \text{Majority}_{j \text{ such that } x_j = 1}\, r^k_j$

$b_k = r^k_i$

$E[\Delta(a,b)] = t/2 + x_i \cdot t^{1/2}$