Outline

1. Information Theory Concepts

2. Distances Between Distributions

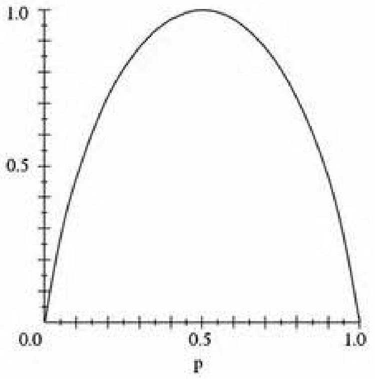
3. An Example Communication Lower Bound – Randomized 1-way Communication Complexity of the INDEX problem

Discrete Distributions

- Consider distributions p over a finite support of size n:
  - p = $(p_1, p_2, p_3, \dots, p_n)$
  - 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

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

(symmetric)

Conditional and Joint Entropy

- Let X and Y be random variables

- (Conditional Entropy)

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

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

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

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

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

Conditioning Cannot Increase Entropy

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

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

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

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) ≤ $H(X \mid Z)$

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’) ≤ $H(P_e)$ + 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

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. Distances Between Distributions

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

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 $\sqrt{p} = (\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_n}),\ \sqrt{q} = (\sqrt{q_1}, \sqrt{q_2}, \dots, \sqrt{q_n})$
  - Note that $\sqrt{p}$ and $\sqrt{q}$ are unit vectors
  - h(p,q) = $\frac{1}{\sqrt{2}}|\sqrt{p}$ - $\sqrt{q}|_2 = \frac{1}{\sqrt{2}}\left(\sum_i\left(\sqrt{p_i} - \sqrt{q_i}\right)^2\right)^{.5}$

- 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

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

$$
\begin{aligned}
&= \frac{1}{2}\left(1 + 1 - 2\,\langle\sqrt{p,q}, \sqrt{p',q'}\,\rangle\right) \\
&= 1 - \textstyle\sum_{i,j}\sqrt{p_i}\sqrt{q_j}\sqrt{p'_i}\sqrt{q'_j} \\
&= 1 - \textstyle\sum_i \sqrt{p_i}\,\sqrt{p'_i} \cdot \sum_j \sqrt{q_j}\ \sqrt{q'_j} \\
&= 1 - (1 - h^2(p,p')) \cdot (1 - h^2(q,q'))
\end{aligned}
$$

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)

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

- (Squared Hellinger Lower Bounded by Squared Variation Distance)

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

- (Variation Distance Upper Bounds Distinguishing Probability)

½ + δ/2

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

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

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

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, $\mathrm{Pr}\left[\mathrm{X}_\mathrm{j}' = \mathrm{X}_\mathrm{j}\right] \geq \frac{2}{3}$

- By Fano's inequality, for all j,

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

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

- Give a_1, …, 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, …, 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, $H(X_i \mid M, X_{< i}) < H(\delta)$ if Bob can predict X_i with probability $> 1- \delta$ from $M, X_{< i}$
- CC_δ(Augmented-Index) $> I(M ; X) \geq n(1-H(\delta))$

Log n Bit Lower Bound for Estimating Norms

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

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

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

- Gap Hamming Problem: Hamming distance $\Delta(x,y) > n/2 + \varepsilon n$ or $\Delta(x,y) < n/2$

- 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 $\{0,1\}^t$

$t = \varepsilon^{-2}$

$x \in \{0,1\}^t$

$\downarrow$

$a \in \{0,1\}^t$

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

$i \in [t]$

$\downarrow$

$b \in \{0,1\}^t$

$b_k = r^k{}_i$

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

1-Way Distributional Communication of Index

- Alice has $x \in \{0,1\}^n$
- Bob has $i \in [n]$
- Alice sends a (randomized) message M to Bob
- $I(M ; X) = \text{sum}_i \, I(M ; X_i \mid X_{< i})$

  ¸ $\text{sum}_i \, I(M; X_i)$

  $= n - \text{sum}_i \, H(X_i \mid M)$
- Fano: $H(X_i \mid M) < H(\delta)$ if Bob can guess X_i with probability $> 1- \delta$
- $CC_\delta(\text{Index}) \geq I(M ; X) \geq n(1-H(\delta))$

The same lower bound applies if the protocol is only correct on average over x and i drawn independently from a uniform distribution

Distributional Communication Complexity

f(X,Y)?

X

Y

- $(X, Y) \sim \mu$

- μ-distributional complexity $D_\mu(f)$: the minimum communication cost of a protocol which outputs f(X,Y) with probability 2/3 for $(X, Y) \sim \mu$
  - Yao's minimax principle: $\mathrm{R(f)} = \max_{\mu} D_\mu(\mathrm{f})$

- 1-way communication: Alice sends a single message M(X) to Bob

Indexing is Universal for Product Distributions [Kremer, Nisan, Ron]

- Communication matrix A_f of a Boolean function $f: \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}$ has (x,y)-th entry equal to f(x,y)

- $\max_{product\ \mu} D_\mu$(f) = Θ(VC-dimension) of A_f

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

- Implies a reduction from Index is optimal for product distributions

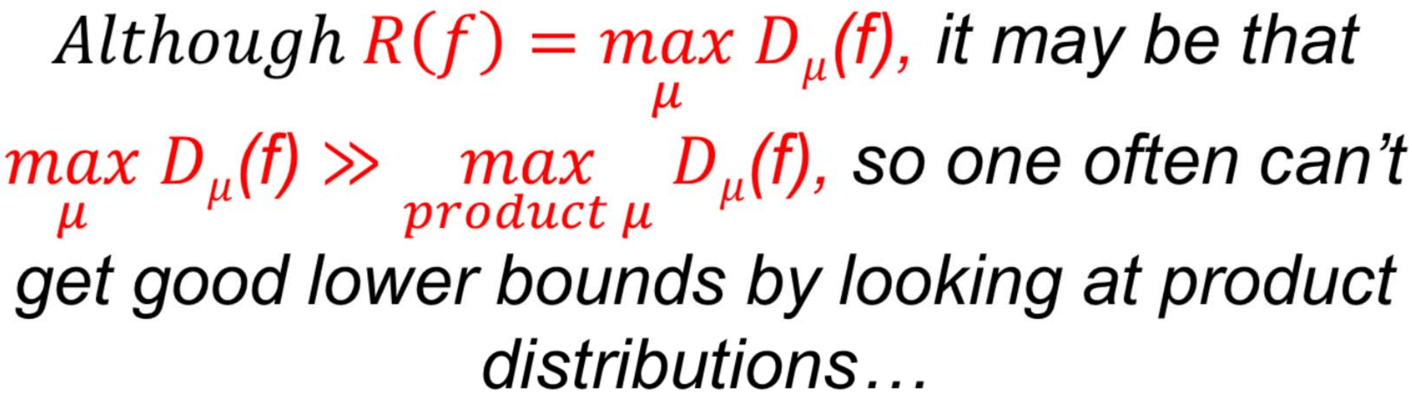
Indexing with Low Error

- Index Problem with 1/3 error probability and 0 error probability both have $\Omega(n)$ communication

- Sometimes, want lower bounds in terms of error probability

- Indexing on Large Alphabets:
  - Alice has $x \in \{0,1\}^{n/\delta}$ with $wt(x) = n$, Bob has $i \in [n/\delta]$
  - Bob wants to decide if $x_i = 1$ with error probability δ
  - [Jayram, W] 1-way communication is $\Omega(n \log(1/\delta))$
  - Can be used to get an $\Omega(\log\left(\frac{1}{\delta}\right))$ bound for norm estimation
  - We've seen an $\Omega(\log n + \epsilon^{-2} + \log\left(\frac{1}{\delta}\right))$ lower bound for norm estimation
  - There is an $\Omega(\epsilon^{-2} \log \frac{1}{\delta} \log n)$ bit lower bound

Beyond Product Distributions



Although $R(f) = \max_{\mu} D_{\mu}(f)$*, it may be that* $\max_{\mu} D_{\mu}(f) \gg \max_{\text{product } \mu} D_{\mu}(f)$*, so one often can't get good lower bounds by looking at product distributions…*

Example: set disjointness

Non-Product Distributions

- Needed for stronger lower bounds
- Example: approximate $|x|_1$ up to a multiplicative factor of B in a stream
  - Lower bounds for p-norms

Gap_∞ (x,y) Problem

$x \in \{0, \ldots, B\}^n$ $y \in \{0, \ldots, B\}^n$

- Promise: $|x-y|_1 \leq 1$ or $|x-y|_1 \geq B$
- Hard distribution non-product
- $\Omega(n/B^2)$ lower bound [Saks, Sun] [Bar-Yossef, Jayram, Kumar, Sivakumar]

Direct Sums

- Gap_∞ (x,y) doesn't have a hard product distribution, but has a hard distribution $\mu = \lambda^n$ in which the coordinate pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are independent
  - w.pr. 1-1/n, (x_i, y_i) random subject to $|x_i - y_i| \leq 1$
  - w.pr. 1/n, (x_i, y_i) random subject to $|x_i - y_i| \geq B$
- Direct Sum: solving Gap_∞(x,y) requires solving n single-coordinate sub-problems g
  - Communication is not additive, but information is!
- In g, Alice and Bob have $J,K \in \{0, \ldots, B\}$, and want to decide if $|J-K| \leq 1$ or $|J-K| \geq B$