Estimating the Number of Non-Zero Entries

• $|x|_0 = |\{i \text{ such that } x_i \neq 0\}|$

• How can we output a number $Z$ with $(1 - \epsilon)Z \leq |x|_0 \leq (1 + \epsilon)Z$ with prob. $9/10$?
  • Want $O((\log n)/\epsilon^2)$ bits of space

• Suppose $|x|_0 = O(\frac{1}{\epsilon^2})$. What can we do in this case?

• Use our algorithm for recovering a k-sparse vector from last time, $k = O\left(\frac{1}{\epsilon^2}\right)$
  • What is another way?

• But what if $|x|_0 \gg \frac{1}{\epsilon^2}$?
Estimating the Number of Non-Zero Entries

- Suppose we somehow had an estimate $Z$ with $Z \leq |x|_0 \leq 2Z$, what could we do?
- Independently sample each coordinate $i$ with probability $p = \frac{100}{(Z \epsilon^2)}$
- Let $Y_i$ be an indicator random variable if coordinate $i$ is sampled
- Let $y$ be the vector restricted to coordinates $i$ for which $Y_i = 1$
- $E[|y|_0] = \sum_{i \text{ such that } x_i \neq 0} E[Y_i] = p|x|_0 \geq \frac{100}{\epsilon^2}$
- $Var[|y|_0] = \sum_{i \text{ such that } x_i \neq 0} Var[Y_i] \leq \frac{200}{\epsilon^2}$
- $Pr\left[||y|_0 - E[|y|_0]| > \frac{100}{\epsilon}\right] \leq \frac{Var[|y|_0] \epsilon^2}{100^2} \leq \frac{1}{50}$
- Use sparse recovery or CountSketch to compute $|y|_0$ exactly
- Output $\frac{|y|_0}{p}$

But we don’t know $Z$...
Estimating the Number of Non-Zero Entries

• Guess $Z$ in powers of 2
• Since $0 \leq |x|_0 \leq n$, there are $O(\log n)$ guesses
• The $i$-th guess $Z = 2^i$ corresponds to sampling each coordinate with probability $p = \min(1, \frac{100}{2^i \epsilon^2})$
• Sample the coordinates as nested subsets $[n] = S_0 \supseteq S_1 \supseteq S_2 \supseteq \cdots S_{\log n}$
• Run previous algorithm for each guess
• One of our guesses $Z$ satisfies $Z \leq |x|_0 \leq 2Z$ and we should use that guess

• But how do we know which one?
Estimating the Number of Non-Zero Entries

- Use the largest guess $Z = 2^i$ for which $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$

- If $\frac{800}{\epsilon^2} \leq E[|y|_0] \leq \frac{1600}{\epsilon^2}$, then $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$ with probability $1 - O(\epsilon^2)$

- If $\frac{100}{\epsilon^2} \leq E[|y|_0] \leq \frac{200}{\epsilon^2}$, then $|y|_0 < \frac{400}{\epsilon^2}$ with probability at least $1 - O(\epsilon^2)$
  - Use nested subset property to conclude this also holds for larger $i$

- So with probability $1 - O(\epsilon^2)$, we choose an $i$ for which $\frac{200}{\epsilon^2} \leq E[|y|_0] \leq \frac{1600}{\epsilon^2}$

- At most 4 such indices $i$, and all 4 of them satisfy $|y|_0 = (1 \pm \epsilon)E[|y|_0]$ simultaneously with probability $1-4/50$. So doesn’t matter which $i$ we choose

- Overall, our success probability is $1 - O(\epsilon^2) - 4/50 > 4/5$
What is Our Overall Space Complexity?

• If we use our k-sparse recovery algorithm for $k = 0 \left(\frac{1}{\epsilon^2}\right)$, then it takes $O\left(\frac{\log n}{\epsilon^2}\right)$ bits of space in each of $\log n$ levels, so $O\left(\frac{\log^2 n}{\epsilon^2}\right)$ total bits of space ignoring random bits
  • How much randomness do we need?
  • Pairwise independence is enough for Chebyshev’s inequality
  • Implement nested sampling by choosing a hash function $h: [n] \rightarrow [n]$, checking if first $i$ bits of $h(j) = 0$
  • $O(\log n)$ bits of space for the randomness

• Can improve to $O\left(\frac{\log n (\log(\frac{1}{\epsilon}) + \log \log n)}{\epsilon^2}\right)$ bits. How?

• Just need to know number of non-zero counters, so reduce counters from $\log n$ bits to $O(\log \left(\frac{1}{\epsilon}\right) + \log \log n)$ bits
Reducing Counter Size

- In $O(\log n)$ sampling levels, we have $O\left(\frac{1}{\epsilon^2}\right)$ counters, each of $O(\log n)$ bits.

- At most $O\left(\frac{\log^2 n}{\epsilon^2}\right)$ prime numbers dividing any of these counters.

- Choose a random prime $q = O\left(\frac{\log^2 n \left(\log \log n + \log\left(\frac{1}{\epsilon}\right)\right)}{\epsilon^2}\right)$. Unlikely that $q$ divides any counter.

- Just maintain our sparse recovery structure mod $q$, so $O\left(\frac{(\log \log n + \log\left(\frac{1}{\epsilon}\right))}{\epsilon^2}\right)$ bits per each of $O(\log n)$ sparse recovery instances.
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 $n$:

  • $p = (p_1, p_2, p_3, \ldots, p_n)$
  
  • $p_i \in [0,1]$ 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) $H(X) = \sum_i p_i \log_2 \left( \frac{1}{p_i} \right)$
  
  • 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)
  
  \[ H(X \mid Y) = \sum_y H(X \mid Y = y) \Pr[Y = y] \]

• (Joint Entropy)
  
  \[ H(X, Y) = \sum_{x,y} \Pr[(X,Y) = (x,y)] \log(1/\Pr[(X,Y) = (x,y)]) \]
Chain Rule for Entropy

• (Chain Rule) \( H(X,Y) = H(X) + H(Y \mid X) \)

• Proof:

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

\[
H(X \mid Y) - H(X) = \sum_{xy} \Pr[Y = y] \Pr[X = x \mid Y = y] \log \left( \frac{1}{\Pr[X=x \mid Y=y]} \right) \\
- \sum_x \Pr[X = x] \log \left( \frac{1}{\Pr[X=x]} \right) \sum_y \Pr[Y = y \mid X = x] \\
= \sum_{x,y} \Pr[X = x, Y = y] \log \left( \frac{\Pr[X=x]}{\Pr[X=x \mid Y=y]} \frac{\Pr[X=x \mid Y=y]}{\Pr[(X,Y) = (x,y)]} \right) \\
\leq \log(\sum_{x,y} \Pr[X = x, Y = y]) \cdot \frac{\Pr[X=x \mid Y=y]}{\Pr[(X,Y) = (x,y)]} \\
= 0
\]

where the inequality follows by Jensen’s inequality.

If \( X \) and \( Y \) are independent \( H(X \mid Y) = H(X) \).
Mutual Information

• (Mutual Information) \( I(X ; Y) = H(X) - H(X | Y) \)
  \[ = H(Y) - H(Y | X) \]
  \[ = I(Y ; X) \]

Note: \( I(X ; X) = H(X) - H(X | X) = H(X) \)

• (Conditional Mutual Information)
  \[ I(X ; Y | Z) = H(X | Z) - H(X | Y, Z) \]

Is \( I(X ; Y | Z) \geq I(X ; Y) \)? Or is \( I(X ; Y | Z) \leq I(X ; Y) \)? Neither!
Mutual Information

• Claim: For certain $X, Y, Z$, we can have $I(X ; Y | Z) \leq I(X ; Y)$

• Consider $X = Y = Z$

• Then,
  • $I(X ; Y | Z) = H(X | Z) - H(X | Y, Z) = 0 - 0 = 0$
  • $I(X ; Y) = H(X) - H(X | Y) = H(X) - 0 = H(X)$

• 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 $X = Y + Z \text{ mod } 2$, where $X$ and $Y$ are uniform in $\{0,1\}$

• Then,
  • $I(X ; Y | Z) = H(X | Z) - H(X | Y, Z) = 1 - 0 = 1$
  • $I(X ; Y) = H(X) - H(X | Y) = 1 - 1 = 0$

• 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, $I(X_1, ..., X_n; Z) = \sum_i I(X_i; Z | X_1, ..., X_{i-1})$
Fano’s Inequality

• For any estimator $X'$: $X \rightarrow Y \rightarrow X'$ with $P_e = \Pr[X' \neq X]$, we have
  $$H(X \mid Y) \leq H(P_e) + P_e \cdot \log(|X| - 1)$$

Here $X \rightarrow Y \rightarrow X'$ 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 \rightarrow Y \rightarrow 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 $X \to Y \to X'$ with $P_e = \Pr[X \neq X']$, we have $H(X \mid Y) \leq H(P_e) + P_e (\log_2 |X| - 1)$.

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

\[
H(E, X \mid X') = H(X \mid X') + H(E \mid X, X') = H(X \mid X')
\]

\[
H(E, X \mid X') = H(E \mid X') + H(X \mid E, X') \leq H(P_e) + H(X \mid E, X')
\]

But $H(X \mid E, X') = \Pr(E = 0)H(X \mid X', E = 0) + \Pr(E = 1)H(X \mid X', E = 1)$

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

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

By Data Processing, $H(X \mid Y) \leq H(X \mid X') \leq H(P_e) + P_e \cdot \log_2 (|X| - 1)$
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 \( I(X ; Y) = H(X) - H(X \mid Y) = 0. \)

• Best predictor \( X' \) of \( X \) is \( X = 1. \)

• \( P_e = \Pr[X' \neq X] = 1 - p_1 \)

• \( H(X \mid Y) \leq H(p_1) + (1 - p_1) \log_2 (n - 1) \) predicted by Fano’s inequality

• But \( H(X) = H(X \mid Y) \) 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 \( (p_1, \frac{1-p_1}{n-1}, ..., \frac{1-p_1}{n-1}) \)

• \( H(X) = \sum_i p_i \log \left( \frac{1}{p_i} \right) \)

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

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

\[ = H(p_1) + (1 - p_1) \log(n - 1) \]
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

- 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 $\mu$ 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 $j$, $\Pr[X_j' = X_j] \geq \frac{2}{3}$

• By Fano’s inequality, for all $j$,
  $$H(X_j | M) \leq H\left(\frac{2}{3}\right) + \frac{1}{3} (\log_2 2 - 1) = H\left(\frac{1}{3}\right)$$
1-Way Communication of Index Continued

• Consider the mutual information $I(M ; X)$

• By the chain rule,

$$I(X ; M) = \Sigma_i I(X_i ; M | X_{<i})$$

$$= \Sigma_i H(X_i | X_{<i}) - H(X_i | M , X_{<i})$$

• Since the coordinates of $X$ are independent bits, $H(X_i | X_{<i}) = H(X_i) = 1$.

• Since conditioning cannot increase entropy,

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

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

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 \( \text{Alg}(s(a)) \) to Bob

2. Bob computes \( \text{Alg}(s(a), s(b)) \)

3. If Bob solves \( g(a,b) \), space complexity of \( \text{Alg} \) 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 $\text{Alg}(s(a), s(b)) = \text{Alg}(s(a)) + 1$ 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
• $I(M ; X) = \sum_i I(M ; X_i | X_{<i})$
  
  
  
  $= n - \sum_i H(X_i | M, X_{<i})$

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

\[ x \in \{0,1\}^n \quad y \in \{0,1\}^n \]

- **Gap Hamming Problem**: Hamming distance \( \Delta(x,y) > n/2 + 2\epsilon n \) or \( \Delta(x,y) < n/2 + \epsilon n \)
- Lower bound of \( \Omega(\epsilon^{-2}) \) for randomized 1-way communication [Indyk, W], [W], [Jayram, Kumar, Sivakumar]
- Gives \( \Omega(\epsilon^{-2}) \) 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 \( \{0,1\}^t \)

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

\( x \in \{0,1\}^t \)

\( a \in \{0,1\}^t \)

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

\( i \in [t] \)

\( b \in \{0,1\}^t \)

\( b_k = r^k_i \)

\[ E[\Delta(a,b)] = t/2 - x_i \cdot t^{1/2} \]
Aspects of 1-Way Communication of Index

• Alice has \( x \in \{0,1\}^n \)
• Bob has \( i \in [n] \)
• Alice sends a (randomized) message \( M \) to Bob
• \( I(M ; X \mid R) = \sum_i I(M ; X_i \mid X_{<i}, R) \)
  \( \geq \sum_i I(M; X_i \mid R) \)
  \( = n – \sum_i H(X_i \mid M, R) \)
• Fano: \( H(X_i \mid M, R) \leq H(\delta) \) if Bob can guess \( X_i \) with probability \( > 1- \delta \)
• \( CC_\delta(\text{Index}) \geq I(M ; X \mid R) \geq n(1-H(\delta)) \)

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

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

- \(\mu\)-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: \(R(f) = \max_\mu D_\mu(f)\)

- 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 $f: X \times Y \rightarrow \{0,1\}$ has $(x,y)$-th entry equal to $f(x,y)$

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

• Implies a reduction from Index is optimal for product distributions

\[
\begin{bmatrix}
0 & 0 & 1 \\
1 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
1 & 1 & 0
\end{bmatrix}
\]
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 $\text{wt}(x) = n$, Bob has $i \in [n/\delta]$
  • Bob wants to decide if $x_i = 1$ with error probability $\delta$
  • [Jayram, W] 1-way communication is $\Omega(n \log(1/\delta))$
  • Can be used to get an $\Omega(\log(1/\delta))$ bound for norm estimation
  • We’ve seen an $\Omega(\log n + \epsilon^{-2} + \log(1/\delta))$ lower bound for norm estimation
  • There is an $\Omega(\epsilon^{-2} \log \frac{1}{\delta} \log n)$ bit lower bound