Adversary

Lower Bounds in Streaming

Piotr Indyk

MIT

Streaming Algorithms

- Norm estimation: $(1+\varepsilon)$-approximation of $||x||_p$, $x \in R^m$, under a sequence of n updates
  - $O(\log(n+m)/\varepsilon^2)$ bits for $p \in (0,2]$

    (excluding randomness)
- Heavy hitters/sparse approximations
- Question: are these algorithms (nearly) optimal ?

Lower Bounds in Streaming

- Two techniques:
  - – Pigeonhole principle: need enough space to distinguish "different" inputs
  - – Communication complexity
- PP is really a special case of CC (but is often easier to apply)
- Today:
  - Randomness and approximation are both necessary for estimating $||x||_0$ in polylog (n+m) space (even in the insertions-only case)
  - Need $\Omega(1/\varepsilon^2)$ bits to $(1+\varepsilon)$-approximate $||x||_2$
  - Need $\Omega(k \log (m/k))$ measurements for the l1/l1 guarantee [Indyk-Khanh-Price'08]

New!

Pigeonhole Principle

Estimating $||x||_0$

- Warmup theorem: any deterministic exact algorithm for computing $||x||_0$ needs $\Omega(m)$ bits of space
- Proof:
  - Assume there is an algorithm A using $M=o(m)$ bits of space
  - Take any vector $y \in \{0,1\}^m$, $||y||_0=m/2$
  - Feed the coordinates of y to A
  - Let $A[y]$ be the state of A at the end of this process, and E be the estimation of $||y||_0$ (i.e., $E= ||y||_0$)
  - We can decode y from $A[y]$:
    - For any $z \in \{0,1\}^m$, $||z||_0=m/2$, feed z to A in state $A[y]$, obtaining $A[y \circ z]$
    - The algorithm computes an estimation E' of $||y+z||_0$ (i.e., $E'= ||y+z||_0$)
    - We have $y=z$ iff $E=E'$
  - Therefore

$$2^M \geq \text{ number of } y\text{'s} = \exp(\Omega(m))$$

Estimating $||x||_0$, ctd.

- Upgraded theorem: any deterministic c-approximate algorithm for computing $||x||_0$ needs $\Omega(m)$ bits of space, for $c<2$
  - Estimation E such that $||x||_0 \leq E < c||x||_0$
- Proof:
  - For any $y \in \{0,1\}^m$, let $ECC(y) \in \{0,1\}^{m'}$, $m'=O(m)$ be such that:
    - $||ECC(y)||_0 = m'/a$, $a=\Theta(1)$
    - For any $y \neq z$, the distance $||ECC(y)-ECC(z)||_0 \geq 2m'(c-1)/a$ (which implies that $||ECC(y+z)||_0 \geq m'/a + m'(c-1)/a = m'c/a$)
  - Take any y
  - Feed the coordinates of $ECC(y)$ to A
  - The remainder of the argument essentially as before (except that $y=z$ iff $E' < m'c/a$)

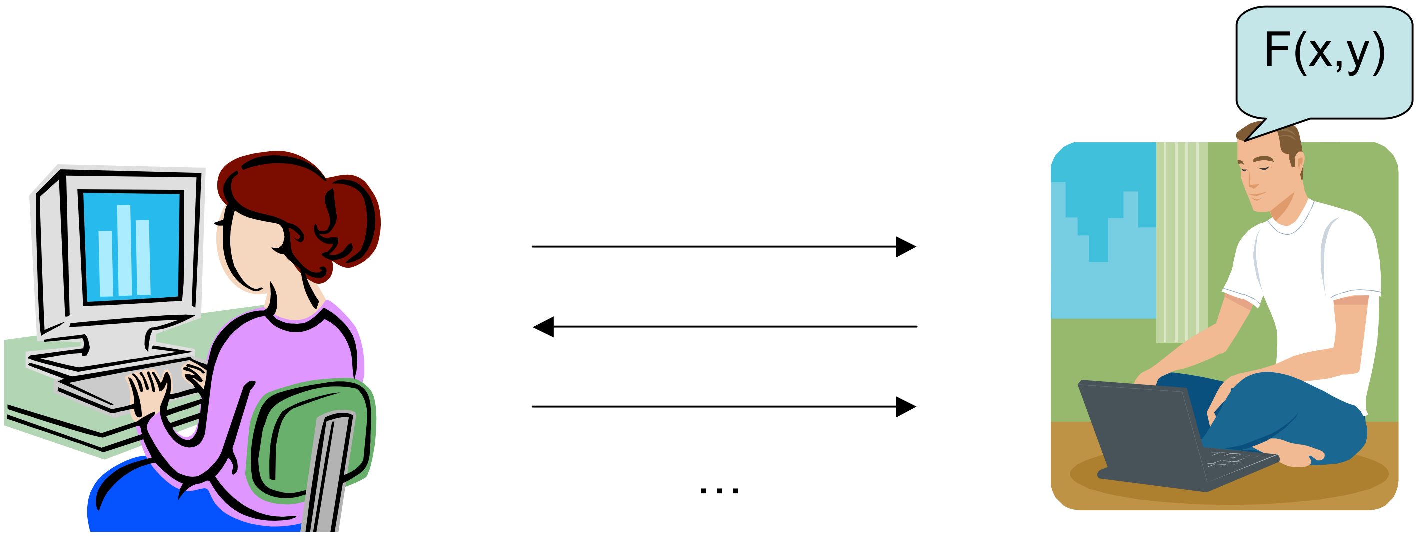
Estimating $||x||_0$, ctd.

- Upgraded theorem 2: any randomized exact algorithm for computing $||x||_0$ needs $\Omega(m)$ bits of space
- Proof:
  - Assume $o(m)$ space, and the probability of error $<1/8$
  - Take any ECC with minimum distance $m'/4$
  - Take any y
  - Feed the coordinates of $ECC(y)$ to A
  - With prob. $1/2$ we can recover z such that $||z-ECC(y)||_0 < m'/4$

    $\rightarrow$ can recover y
    - In parallel, for any $i=1..m'$, feed e_i to A with state $A[ECC(y)]$, obtaining estimate E_i
    - Set $z_i=0$ iff $E_i > m/a'$ (works correctly with prob. $1-1/8$)
    - Markov inequality implies that the fraction of errors is $<1/4$ with prob. $1/2$
  - There is a choice of random coin tosses which works for half of the vectors y (i.e., for those y's, we can recover y from $A[ECC(y)]$)
  - The rest of the argument as before

Communication Complexity

Communication Complexity



Alice: $x \in \{0,1\}^m$ Bob: $y \in \{0,1\}^m$

- Resources:
  - # bits
  - # rounds
    - Today, we will be only interested in one-round protocols
- Probability of error: some constant $\delta>0$
- See [Kushilevitz-Nisan] for more
  (and there is much more)

Indexing

- (Balanced) indexing problem:
  – Alice: a vector $x \in \{0,1\}^m$, $||x||_0 = m/2$
  – Bob: an index $i=1\ldots m$
  – Goal: compute $f(x,i)=x_i$
- Theorem: any randomized one-round protocol for indexing has $\Omega(m)$ bit complexity
- Proof: pigeonhole principle as applied earlier

Gap Dot Product

- (Gap) parameter Δ
- Alice: a vector $u \in R^m$, $||u||_2=1$ (with $O(\log m)$ bits)
- Bob: a vector $v \in R^m$, $||v||_2=1$
- Goal:
  - If $u*v=0$, return 0
  - If $u*v \geq \Delta$, return 1
- Theorem: the randomized one-round CC of GDP with gap $\Delta=1/(m/2)^{1/2}$ is $\Omega(m)$
- Proof: via reduction from indexing:
  - Alice: computes $u= \Delta x$
  - Bob: computes $v=e_i$
  - Fact: $u*v= \Delta x_i$

Space complexity of L_2 norm estimation

- Theorem: any streaming algorithm for estimating the L_2 norm of an m-dimensional vector x up to a factor of $1 \pm \Delta$, $\Delta = c/m^{1/2}$, requires $\Omega(m)$ bits for some constant $c>0$

  (even if coordinates of x have $O(\log m)$ bits)
- Proof:
  - Assume we have an M-space streaming algorithm that computes $(1 \pm \Delta)||x||_2$
  - Then we have an M-space streaming algorithm that, given a stream $u \circ v$, $||u||_2 = ||v||_2 = 1$, computes $u^* v \pm O(\Delta)$ (Lecture 4)
    - Using the equality $||u-v||_2^2 = ||u||_2^2 + ||v||_2^2 - 2u^* v$
  - Then we have an M-bit one-round protocol that solves GDP with gap $1/(m/2)^{1/2}$ (assuming c small enough)
  - Ergo, $M = \Omega(m)$

Back to Pigeonhole Principle

Lower bound for l1/l1

- Compressive sensing setup: want an $M \mathrm{x} m$ sketch matrix A such that:
  - Given: Ax for an arbitrary vector x
  - Can obtain: an approximation x^* such that

$$||x^*-x||_1 \leq C\, Err_1^k(x)$$

  where $Err_1^k(x)=\min_{x'} ||x'-x||_1$ over all x' that are k-sparse
- Will show that $M= \Omega(k \log(m/k))$

  (if C is an absolute constant)

Error-correcting code

- Let $E \subseteq \{0,1\}^m$ be a set of k-sparse vectors such that for any distinct $y1,y2 \in E$ we have

$$\|y1-y2\|_1 > k$$

- We can have $|E|>\exp(c\ k \log (m/k))$ for some absolute constant c
- Define $r=k/(2C+2)$
- We consider signals $x = y + z$ where $y \in E$ and $\|z\|_1 \leq r$
  – Clearly, $Err_1^k(x) \leq \|z\|_1 \leq r$

Distinctness

- Lemma: For any $x1=y1+z1$ and $x2=y2+z2$ as in the earlier slide, we have

$$Ax1 \neq Ax2$$

- Proof:
  - Suppose we have $Ax1 = Ax2$
  - We know:
    - Given $Ax1$, our algorithm decodes $x1^*$ s.t. $\|x1-x1^*\|_1 \leq Cr$
    - Given $Ax2$, our algorithm decodes $x2^*$ s.t. $\|x2-x2^*\|_1 \leq Cr$
  - But if $Ax1 = Ax2$ then $x1^* = x2^*$
  - This would imply $\|x1-x2\|_1 \leq 2Cr$
  - Therefore $\|y1-y2\|_1 \leq 2Cr+2r=k$ - a contradiction
- Corollary: Let $B=B_1(0,r)$. Then for any distinct $y1,y2 \in E$ the “affine balls” $A(y1+B)$ and $A(y2+B)$ are disjoint

Pigeonhole

- All "affine balls" $A(y1+B)$ and $A(y2+B)$ are disjoint
- At the same time, for all $y \in E$ we have

  $$y+B \subseteq BB = B_1(0,R),$$

  where $R=k+r = (2C+3)r$
- Therefore, $A(y+B) \subseteq A(BB)$, so $vol(A(BB)) \geq |E|\ vol(A(B))$
- Altogether

  $$\exp(c\ k \log(m/k)) < |E| \leq vol(A(BB)) / vol(A(B)) \leq (2C+3)^M$$
- After applying logarithm on both sides we get

  $$M = \Omega(k \log(m/k))$$