Sampling Algorithms: Lower Bounds and Applications

Ziv Bar-Yossef
Computer Science Division
U.C. Berkeley
http://www.cs.berkeley.edu/~zivi

Ravi Kumar
D. Sivakumar
IBM Almaden Research Center

Sampling Motivation: Computing over Massive Data Sets

Huge collections of raw data:

- Web
- IP packets
- AT&T telephone calls
- WalMart transactions
- Hubble observations

Computational challenges:

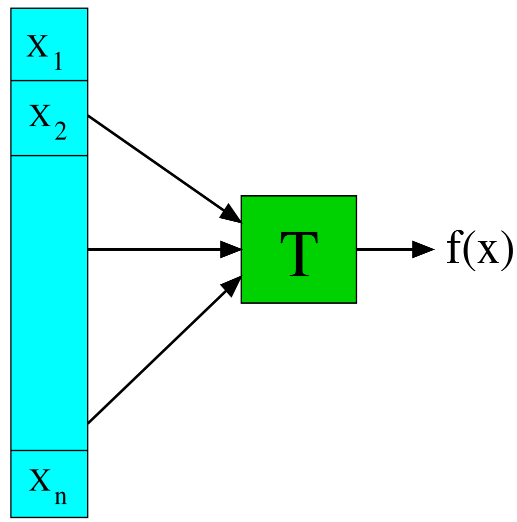
- Run in sub-linear space
- Run in sub-linear time
- Minimize random access to the data

Useful technique:

Query only small (possibly random) portions of the data

Sampling Algorithms

$f : A^n \rightarrow B$	a function of the data set to be computed
$x = x_1, \ldots, x_n \in A^n$	an input data set
T	a sampling algorithm for f



- Query input at random locations
- Can choose query distribution
- Can query adaptively

Sampling Algorithms (cont.)

Worst-case/Expected query complexity:

$S^W(T,x)$	max $\#$ of queries of T on x	(over coin tosses)
$S^E(T,x)$	expected $\#$ of queries of T on x	(over coin tosses)
$S(T)$	max $S(T,x)$	(over inputs x)
$S(f)$	min $S(T)$	(among all T for f)

Examples

- Parameter estimation
- PAC [Valiant 84] and statistical learning [Vapnik 98]
- Property testing [Goldreich et al. 1996]

Our Setting

Goal:

Study the query complexity of general functions $f : A^n \to B$

- arbitrary domain and range
- approximations of f, rather than exact computations
- different types of approximation

Examples

- Low query complexity
  - estimating the mean of n numbers in the interval $[0, 1]$
  - estimating the median of n numbers
- High query complexity
  - estimating the number of distinct elements in a list
  - finding the parity of n bits

Previous Work – Computer Science

- Query complexity of exact computations of Boolean functions cf. [Buhrman and de Wolf 1999]
- Specific lower bounds:
  - Additive approximation of the mean [Canetti et al. 1995]
  - Relative approximation of the mean [Dagum et al. 1995, Schulman and Vazirani 1999]
  - Ratio approximation of the number of distinct elements [Charikar et al. 2000]
  - Median approximation in the quantum model [Nayak and Wu, 1999]

Previous Work – Statistics and Learning

Related Areas:

- Statistical decision theory
- Statistical estimation theory
- Statistical sequential analysis
- Statistical learning theory

Related lower bounds:

- The Crámer-Rao inequality
- VC dimension based sampling lower bounds
- Optimality of the sequential probability ratio test

All hold only for i.i.d. queries.

Uniform Approximation Notion

Definition

An ϵ-approximation of $f : A^n \to B$ is a function $C_{f,\epsilon} : A^n \to 2^B$
$C_{f,\epsilon}(x) =$ "values that are an ϵ-approximation of $f(x)$".

An approximation of f is a family $\{C_{f,\epsilon} : A^n \to 2^B\}_{\epsilon \geq 0}$ that satisfies:

1. $\forall x \in A^n,\ C_{f,0}(x) = \{f(x)\}$ (containment of truth)
2. $\forall x \in A^n, \forall \epsilon' > \epsilon \geq 0,\ C_{f,\epsilon}(x) \subseteq C_{f,\epsilon'}(x)$ (nesting)

Examples ($B = \mathbf{R}$)

Additive $C_{f,\epsilon}(x) = [f(x) - \epsilon, f(x) + \epsilon]$

Relative $C_{f,\epsilon}(x) = [(1-\epsilon)f(x), (1+\epsilon)f(x)]$

Quantile $C_{\text{select}_{q},\epsilon}(x_1, \ldots, x_n) = \{x_i \mid \text{quant}(x_i) \in [q-\epsilon, q+\epsilon]\}$

Randomized Approximation

Definition

Let $\epsilon \geq 0$ (error parameter)

Let $0 \leq \delta < 1$ (confidence parameter)

A sampling algorithm T is said to (ϵ, δ)-approximate f, if for every $x \in A^n$,

$$\Pr\left[T(x) \in C_{f,\epsilon}(x)\right] \geq 1 - \delta$$

Main Results

Two lower bound techniques for the query complexity of general function approximation:

1. Lower bound via block sensitivity
   - bounds the expected query complexity of every function
   - tends to be weak
   - based on a generalization of Boolean block sensitivity [Nisan, 1989]
2. Lower bound via Hellinger distance
   - bounds the worst-case/expected query complexity of symmetric functions
   - tends to be tight
   - based on a reduction to statistical hypothesis testing

Main Idea

Definition

x and y are ϵ-disjoint if $C_{f,\epsilon}(x) \cap C_{f,\epsilon}(y) = \emptyset$.

Observation

Every sampling algorithm that approximates f "distinguishes" between every two ϵ-disjoint inputs.

"Distinguishes" – output distribution when x is the input is far from output distribution when y is the input.

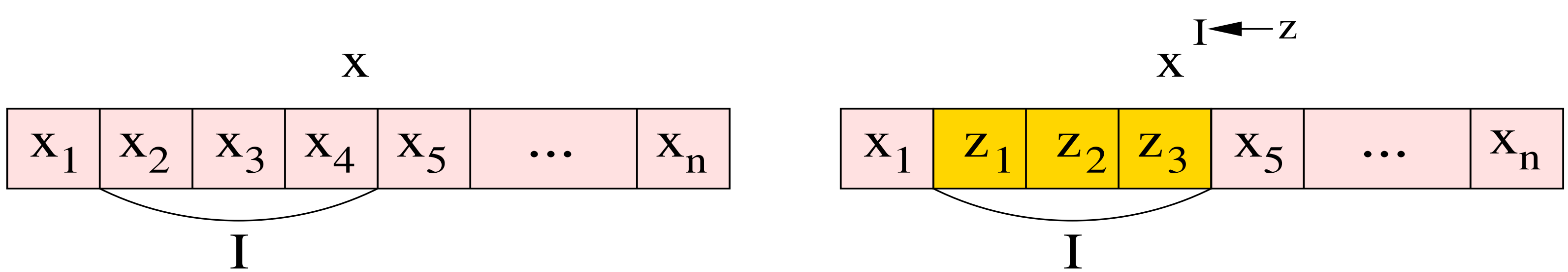
Sampling Lower Bound via Block Sensitivity

Theorem 1

For every f, $S^E(f) \geq (1-2\delta) bs_\epsilon(f)$.

Definitions

f is ϵ-sensitive to block $I \subseteq [n]$ on input x, if $\exists z \in A^{|I|}$ s.t. x and $x^{I \leftarrow z}$ are ϵ-disjoint.



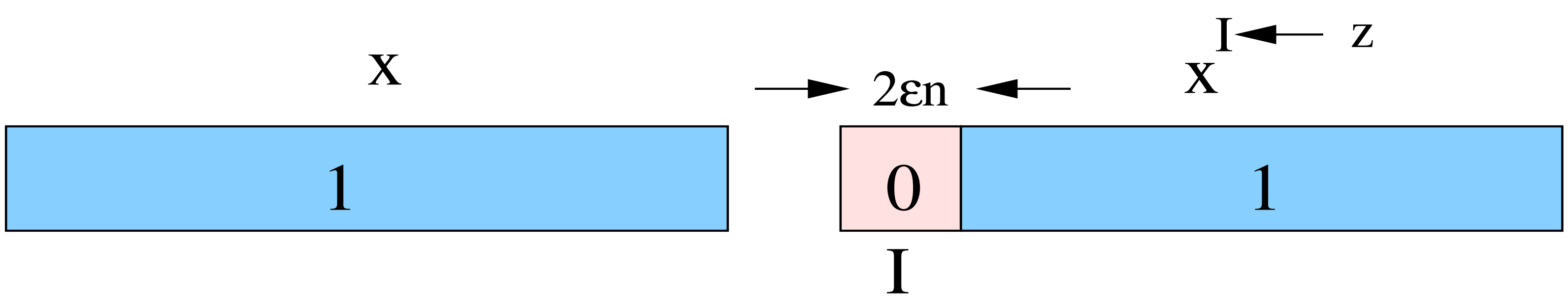
$bs_\epsilon(f,x)$ max # of disjoint blocks to which f is ϵ-sensitive on x

$bs_\epsilon(f)$ $\max\{bs_\epsilon(f,x) \mid x \in A^n\}$

Application: Mean

- mean:
  $\mu_1 : [0,1]^n \to [0,1], \quad \mu_1(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$
- additive approximation
- Best upper bound: $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$

Theorem $S^E(\mu_1) = \Omega\left(\frac{1}{\epsilon}(1-2\delta)\right)$



- $C_{\mu_1,\epsilon}(x) = (1-\epsilon, 1), \quad C_{\mu_1,\epsilon}(x^{I\leftarrow z}) = (1-3\epsilon, 1-\epsilon)$
- $bs_\epsilon(\mu_1, x) = n/|I| = 1/(2\epsilon)$

Sampling Lower Bound for Symmetric Functions

Theorem 2

For any ϵ-symmetric f,

(1) if $S^W(f) \le \sqrt{2\delta n}$,

$$S^W(f) \ge \Omega\left(\frac{1}{h_\epsilon^2(f)} \log \frac{1}{\delta}\right)$$

(2) if $S^W(f) \le n/2 - o(n)$,

$$S^W(f) \ge \Omega\left(\frac{1}{d_\epsilon(f)} \log \frac{1}{\delta}\right)$$

(3) For any f, $\quad S^E(f) \ge S^W(f) / \log \frac{1}{\delta}$

Sampling Lower Bound for Symmetric Functions (cont.)

Definitions

- f is ϵ-symmetric if $\forall x \in A^n, \forall \pi \in S_n$, $C_{f,\epsilon}(x) = C_{f,\epsilon}(\pi(x))$
- U_x = distribution on A induced by:

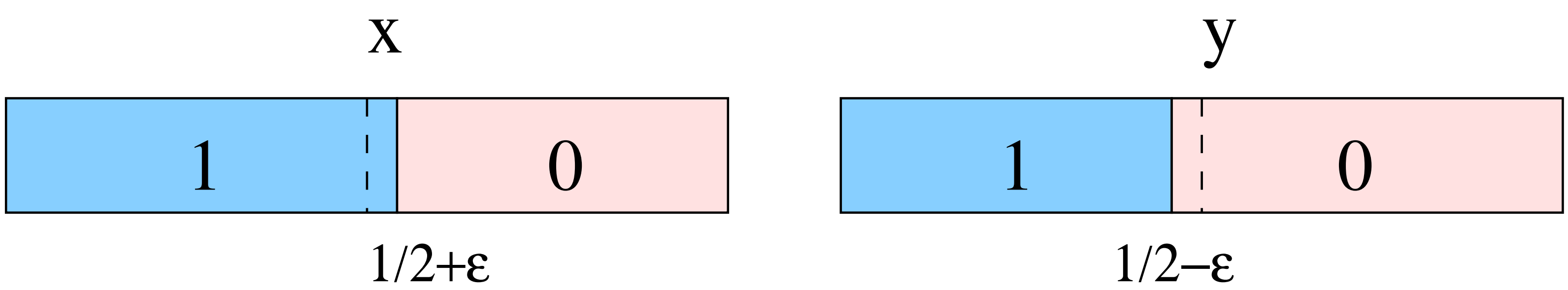
  pick $i \in [n]$ u.a.r., output x_i

- Hellinger distance of two distributions U_x, U_y on A:
  $h^2(U_x, U_y) = \frac{1}{2} \sum_{a \in A} (\sqrt{U_x(a)} - \sqrt{U_y(a)})^2$
- ϵ-minimum Hellinger distance of f:
  $h_\epsilon(f) = \min\{h(U_x, U_y) \mid x, y$ are ϵ-disjoint$\}$

Application: Additive Approximation of the Mean

Theorem (Originally proved in [Canetti et al. 1995])

$$S^W(\mu_1) = \Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$$



- $C_{\mu_1,\epsilon}(x) = (1/2, 1/2 + 2\epsilon), \quad C_{\mu_1,\epsilon}(y) = (1/2 - 2\epsilon, 1/2)$
- $h^2(U_x, U_y) = 1 - 2\sqrt{(1/2 + \epsilon)(1/2 - \epsilon)} \leq 4\epsilon^2$

- Implies extractor and disperser seed length lower bounds (originally, [Radhakrishnan and Ta-Shma 1997])
- Can be generalized to any statistical moment

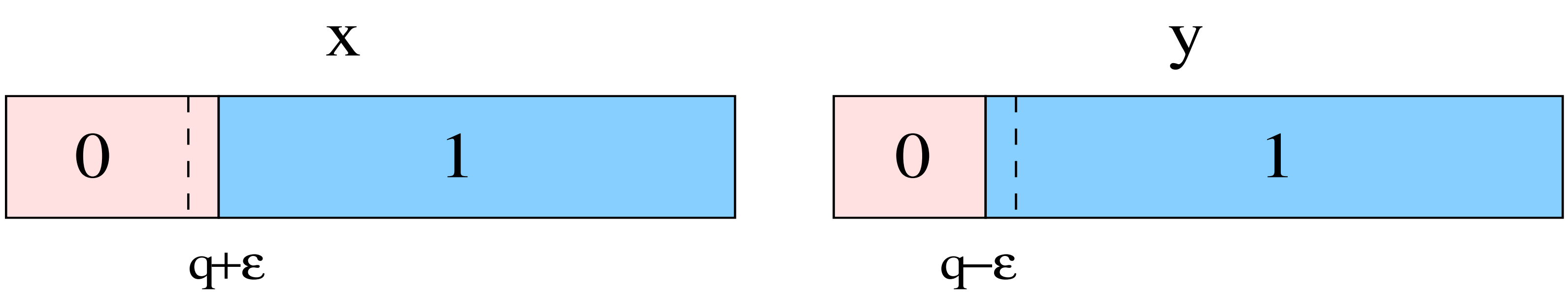
Application: Selection Functions

- Selection function of order q ($0 \leq q \leq 1$):
  $\mathrm{select}_q(x_1, \ldots, x_n) =$ element of relative rank q
- quantile approximation

Theorem

$S^W(\mathrm{select}_q) = \Theta\left(\frac{q(1-q)}{\epsilon^2} \log \frac{1}{\delta}\right)$ for $\epsilon \leq q \leq 1 - \epsilon$

$S^W(\mathrm{select}_q) = \Theta\left(\frac{1}{\epsilon} \log \frac{1}{\delta}\right)$ otherwise



- $C_{\mathrm{select}_{q,\epsilon}}(x) = \{0\}, \; C_{\mathrm{select}_{q,\epsilon}}(y) = \{1\}$
- $h^2(U_x, U_y) \leq \epsilon^2/(q(1-q))$

Application: Frequency Moments

- k-th frequency moment:
  $F_k : [m]^n \to \mathbf{R}, \quad F_k(x_1, \ldots, x_n) = \sum_{j=1}^{m} f_j^k,$
  where f_j is the number of times j appears in $x_1, \ldots, x_n$
- relative approximation

Theorem

$S^W(F_k) \geq \Omega\left(\frac{n^{1-1/k}}{\epsilon^{1/k}} \log \frac{1}{\delta}\right)$ for $k \geq 2$

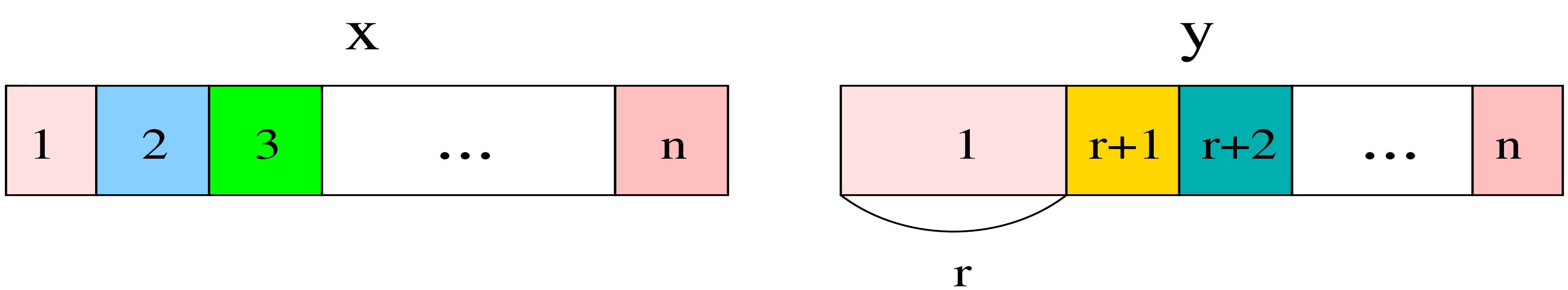
$S^W(F_k) \leq O\left(\frac{kn^{1-1/k}}{\epsilon^{2/(k-1)}} \log \frac{1}{\delta}\right)$ for $k \geq 2$

$S^W(F_0) \geq \Omega\left(\epsilon^2 n \log \frac{1}{\delta}\right)$ (originally, [Charikar et al. 2000])

Application: Frequency Moments (cont.)

Proof

Lower bound for $k \geq 2$:



- $F_k(x) = n, \quad F_k(y) = r^k + (n - r)$
- Choose $r = (3\epsilon n)^{1/k}$
- $d(U_x, U_y) = r/n$

Sampling Lower Bound for Symmetric Functions

Theorem 2

For any ϵ-symmetric f with $S^W(f) \leq O(\sqrt{n})$,

$$S^W(f) \geq \Omega\left(\frac{1}{h_\epsilon^2(f)} \log \frac{1}{\delta}\right)$$

Proof Outline

1. For symmetric functions, wlog, all queries are uniform without replacement.

2. If $k \leq O(\sqrt{n})$, can assume further all queries are uniform with replacement.

3. Fix any two ϵ-disjoint inputs x, y.
   (ϵ, δ)-approximation of f with k queries $\;\Rightarrow\;$ $d(U_x^k, U_y^k) \geq 1 - 2\delta$

4. $d(U_x^k, U_y^k) \geq 1 - 2\delta \;\Rightarrow\; h^2(U_x^k, U_y^k) \geq 1 - 4\delta$

5. $h^2(U_x^k, U_y^k) \geq 1 - 4\delta \;\wedge\; h^2(U_x, U_y) \leq v \;\Rightarrow\; k \geq \Omega(\frac{1}{v} \log \frac{1}{\delta})$

Uniform Queries for Symmetric Functions

Lemma 1
For symmetric functions, wlog, all queries are uniform without replacement.

Proof
T = any sampling algorithm for f.
WLOG, T never queries the same location twice.

Simulate T with uniform queries:

1. Given input x, pick a random permutation $\pi \in S_n$.
2. Simulate T on $y = \pi(x)$.

- Since f is symmetric, any ϵ-approximation of $f(y)$ is also an ϵ-approximation of $f(x)$.
- Queries to x are uniform without replacement.

Uniform Queries with Replacement

Lemma 2

For any symmetric function f with $S^W(f) \leq \sqrt{2\delta n}$, wlog, all queries are uniform with replacement.

Proof

Lemma 1 $+$ a birthday argument

Sampling Algorithm $\Longrightarrow$ Statistical Test

From now on, fix a pair of ϵ-disjoint inputs x, y.

Lemma 3

Let T be a sampling algorithm that (ϵ, δ)-approximates f with k queries. Then, $d(U_x^k, U_y^k) \geq 1 - 2\delta$.

Proof

- By Lemma 2, T "sees" a sample from U_z^k when z is the input.
- Can use T to test whether $z = x$ or $z = y$ with advantage $1 - 2\delta$:
  - Decide "x", if T outputs an ϵ-approximation of $f(x)$.
  - Decide "y" otherwise
- Therefore, variation distance is $\geq 1 - 2\delta$.

Sample Lower Bound for Statistical Tests

State of affairs

- Know $d(U_x^k, U_y^k) \geq 1 - 2\delta$
- Can easily evaluate $d(U_x, U_y)$
- Want lower bound on k

Lemma 4 (cf. [Le Cam, Lo Yang, 1990])

1. $h^2(U_x, U_y) \leq d(U_x, U_y) \leq h(U_x, U_y)\sqrt{2 - h^2(U_x, U_y)}$
2. $1 - h^2(U_x^k, U_y^k) = (1 - h^2(U_x, U_y))^k$

Corollary

1. $d(U_x^k, U_y^k) \geq 1 - 2\delta \quad \Rightarrow \quad h^2(U_x^k, U_y^k) \geq 1 - 4\delta$
2. $h^2(U_x^k, U_y^k) \geq 1 - 4\delta \quad \Rightarrow \quad k \geq \Omega\left(\frac{1}{h^2(U_x, U_y)} \log \frac{1}{\delta}\right)$

Tightness of Bounds

Theorem

For every ϵ-symmetric function f,

$$\frac{1}{h_\epsilon^2(f)} \geq \frac{1}{d_\epsilon(f)} \geq bs_\epsilon(f)$$

Theorem

For every ϵ-symmetric function $f : A^n \to B$,

$$S^W(f) \leq O\left(|A| \frac{1}{d_\epsilon^2(f)} \log \frac{1}{\delta}\right)$$

Conclusions

- Uniform notion of approximation
- Two lower bound techniques for query complexity of general function approximation
- Techniques give “templates” for proving lower bounds
- New lower and upper bounds for the frequency moments
- Unified, simplified versions of several known lower bounds

Open Issues

- Find a tight lower bound for the expected query complexity of symmetric functions.
- Improve the dependence of the Hellinger bound on the number of pairwise disjoint inputs.
- Find tighter lower bounds for non-symmetric functions.

Thank You!