Lecture 13-14: Sublinear-Time Algorithms

John Sylvester Nicolás Rivera Luca Zanetti Thomas Sauerwald

Lent 2019

UNIVERSITY OF CAMBRIDGE

Outline

Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

Usually these algorithms are randomised!

Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

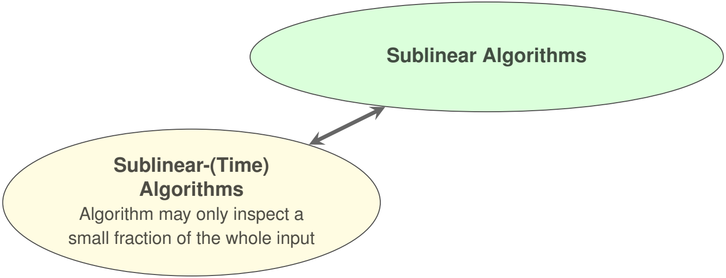
Usually these algorithms are randomised!

Sublinear Algorithms

Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

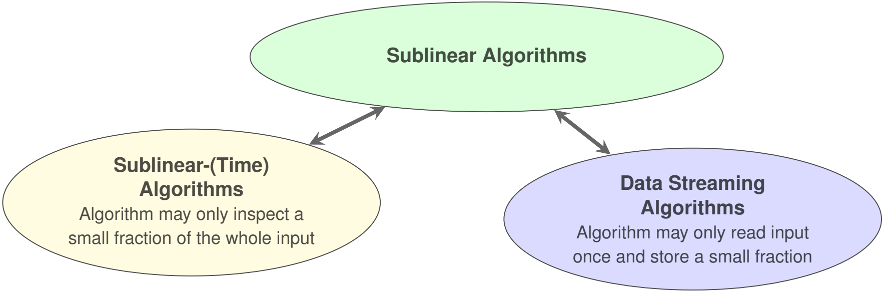
Usually these algorithms are randomised!



Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

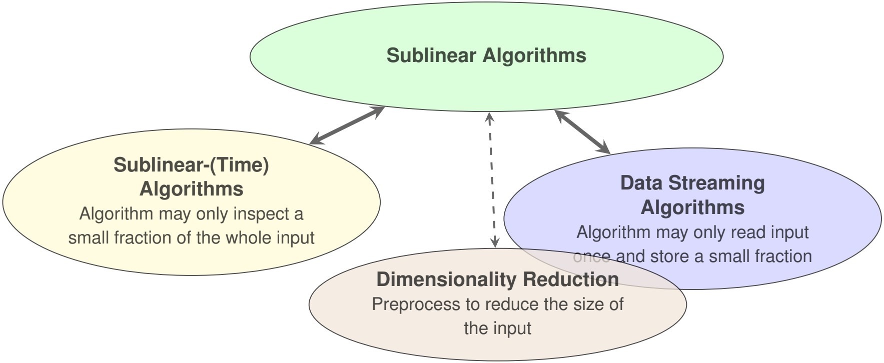
Usually these algorithms are randomised!



Sublinear Algorithms Overview

Sublinear Algorithms: Algorithms that return reasonably good approximate answers without scanning or storing the entire input

Usually these algorithms are randomised!



Motivation

Goal: Estimate properties of **big** probability distributions

Motivation

Goal: Estimate properties of **big** probability distributions

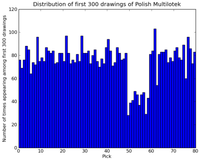
big means that the domain of the finite probability distribution is very large!

Motivation

Goal: Estimate properties of **big** probability distributions

big means that the domain of the finite probability distribution is very large!

- Lottery (are numbers equally likely?)
- Birthday Distribution (is the birthday distribution uniform over 365 days?)



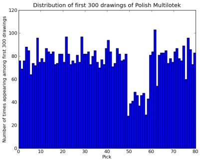
Thanks to Krzysztof Onak (pointer) and Eric Price (graph)

Motivation

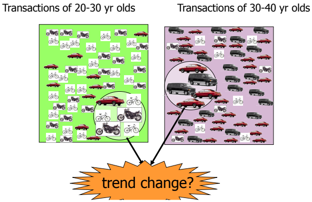
Goal: Estimate properties of **big** probability distributions

big means that the domain of the finite probability distribution is very large!

- Lottery (are numbers equally likely?)
- Birthday Distribution (is the birthday distribution uniform over 365 days?)
- Shopping patterns (are distributions the same or different?)



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)



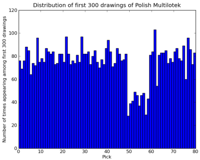
Source: Slides by Ronitt Rubinfeld

Motivation

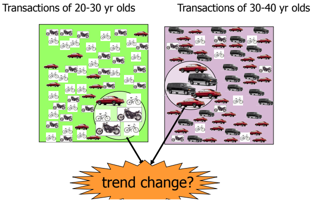
Goal: Estimate properties of **big** probability distributions

big means that the domain of the finite probability distribution is very large!

- Lottery (are numbers equally likely?)
- Birthday Distribution (is the birthday distribution uniform over 365 days?)
- Shopping patterns (are distributions the same or different?)
- Physical Experiment (is the observed distribution close to the prediction?)



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)



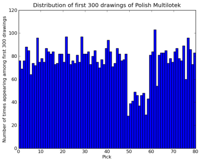
Source: Slides by Ronitt Rubinfeld

Motivation

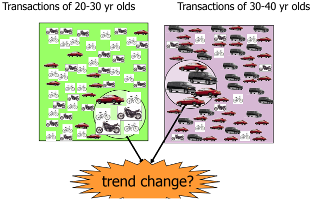
Goal: Estimate properties of **big** probability distributions

big means that the domain of the finite probability distribution is very large!

- Lottery (are numbers equally likely?)
- Birthday Distribution (is the birthday distribution uniform over 365 days?)
- Shopping patterns (are distributions the same or different?)
- Physical Experiment (is the observed distribution close to the prediction?)
- Health (are there correlations between zip code and health condition?)



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)



Source: Slides by Ronitt Rubinfeld

Testing Probability Distribution (Formal Model)

Model

- Given one (or more) probability distribution $p = (p_1, p_2, \ldots, p_n)$
- distribution(s) are unknown, but can obtain independent samples
- also known: n (or a good estimate of it)

Testing Probability Distribution (Formal Model)

Model

- Given one (or more) probability distribution $p = (p_1, p_2, \ldots, p_n)$
- distribution(s) are unknown, but can obtain independent samples
- also known: n (or a good estimate of it)

Cost: number of samples (queries)

Testing Probability Distribution (Formal Model)

Model

- Given one (or more) probability distribution $p = (p_1, p_2, \dots, p_n)$
- distribution(s) are unknown, but can obtain independent samples
- also known: n (or a good estimate of it)

Cost: number of samples (queries)

Questions:

1. Is the distribution p close to the uniform distribution u?
2. Is the distribution p close to some other distribution q?
3. What is $\max_{1 \leq i \leq n} p_i$ (heavy hitter)?
4. Are the distributions p and q independent? ...

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^{n} |p_i - q_i| \in [0, 1]$.

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^{n} |p_i - q_i| \in [0, 1]$.

Examples:

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^{n} |p_i - q_i| \in [0, 1]$.

Examples:

1. $p = (1, 0, \ldots, 0)$, $q = (0, 1, 0, \ldots, 0)$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2}$ and $\|p - q\|_\infty = 1$.

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^{n} |p_i - q_i| \in [0, 1]$.

Examples:

1. $p = (1, 0, \ldots, 0)$, $q = (0, 1, 0, \ldots, 0)$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2}$ and $\|p - q\|_\infty = 1$.
2. $p = (1, 0, \ldots, 0)$, $q = (1/n, 1/n, \ldots, 1/n)$. Then $\|p - q\|_1 = 2 - 2/n$, $\|p - q\|_2 = \sqrt{1 \cdot (1 - 1/n)^2 + (n - 1) \cdot (1/n)^2} = \sqrt{1 - 1/n}$ and $\|p - q\|_\infty = 1 - 1/n$.

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^{n} |p_i - q_i| \in [0, 1]$.

Examples:

1. $p = (1, 0, \ldots, 0)$, $q = (0, 1, 0, \ldots, 0)$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2}$ and $\|p - q\|_\infty = 1$.
2. $p = (1, 0, \ldots, 0)$, $q = (1/n, 1/n, \ldots, 1/n)$. Then $\|p - q\|_1 = 2 - 2/n$, $\|p - q\|_2 = \sqrt{1 \cdot (1 - 1/n)^2 + (n-1) \cdot (1/n)^2} = \sqrt{1 - 1/n}$ and $\|p - q\|_\infty = 1 - 1/n$.
3. $p = (\underbrace{2/n, \ldots, 2/n}_{n/2 \text{ times}}, 0, \ldots, 0)$ and $q = (0, \ldots, 0, \underbrace{2/n, \ldots, 2/n}_{n/2 \text{ times}})$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2 \cdot (n/2) \cdot (2/n)^2} = \sqrt{4/n}$ and $\|p - q\|_\infty = 2/n$.

Testing Uniformity

Testing Uniformity: Is the distribution p close to the uniform distribution u?

Distance between Discrete Distributions

Let p and q be any two distributions over $\{1, 2, \ldots, n\}$. Then:

1. L_1-distance: $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i| \in [0, 2]$,
2. L_2-distance: $\|p - q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \in [0, \sqrt{2}]$,
3. L_∞-distance: $\|p - q\|_\infty = \max_{i=1}^n |p_i - q_i| \in [0, 1]$.

Examples:

1. $p = (1, 0, \ldots, 0)$, $q = (0, 1, 0, \ldots, 0)$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2}$ and $\|p - q\|_\infty = 1$.
2. $p = (1, 0, \ldots, 0)$, $q = (1/n, 1/n, \ldots, 1/n)$. Then $\|p - q\|_1 = 2 - 2/n$, $\|p - q\|_2 = \sqrt{1 \cdot (1 - 1/n)^2 + (n-1) \cdot (1/n)^2} = \sqrt{1 - 1/n}$ and $\|p - q\|_\infty = 1 - 1/n$.
3. $p = (\underbrace{2/n, \ldots, 2/n}_{n/2 \text{ times}}, 0, \ldots, 0)$ and $q = (0, \ldots, 0, \underbrace{2/n, \ldots, 2/n}_{n/2 \text{ times}})$. Then $\|p - q\|_1 = 2$, $\|p - q\|_2 = \sqrt{2 \cdot (n/2) \cdot (2/n)^2} = \sqrt{4/n}$ and $\|p - q\|_\infty = 2/n$.

Disjoint distributions, yet L_2 and L_∞ distances are small!

Outline

Testing Uniformity in the L_1-distance

Objective

Find an efficient tester such that

- Given any probability distribution p and $\epsilon \in (0, 1)$

Testing Uniformity in the L_1-distance

Objective

Find an efficient tester such that

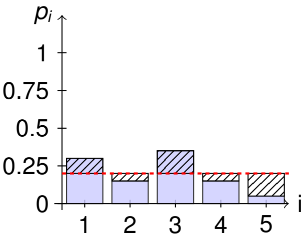
- Given any probability distribution p and $\epsilon \in (0, 1)$
  - If p is the uniform distribution, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq 2/3$,

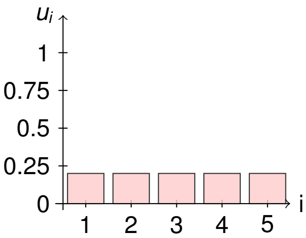
Testing Uniformity in the L_1-distance

Objective

Find an efficient tester such that

- Given any probability distribution p and $\epsilon \in (0,1)$
  - If p is the uniform distribution, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq 2/3$,
  - If p is ϵ-far from uniform ($\sum_{i=1}^{n} |p_i - 1/n| \geq \epsilon$), then $\mathbf{P}[\,\text{REJECT}\,] \geq 2/3$.





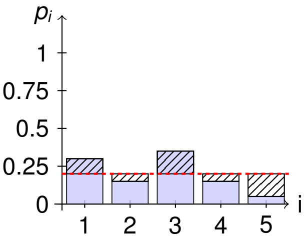
Testing Uniformity in the L_1-distance

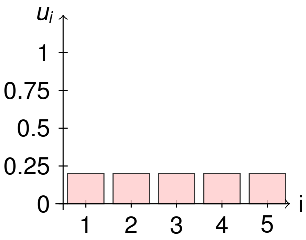
Objective

Find an efficient tester such that

- Given any probability distribution p and $\epsilon \in (0, 1)$
  - If p is the uniform distribution, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq 2/3$,
  - If p is ϵ-far from uniform ($\sum_{i=1}^{n} |p_i - 1/n| \geq \epsilon$), then $\mathbf{P}[\,\text{REJECT}\,] \geq 2/3$.

- tester efficient (sub-linear) $\rightsquigarrow$ different from standard statistical tests!
- tester is allowed to have two-sided error
- there is a "grey area" when p is different from but close to uniform, where the tester may give any result





High Level Idea

Recall: L_1-distance is

$$\sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right|$$

High Level Idea

Recall: L_1-distance is

$$\sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right|$$

First Idea might be to approximate each $p_i - \frac{1}{n}$, but this takes at least $\Omega(n)$ queries.

High Level Idea

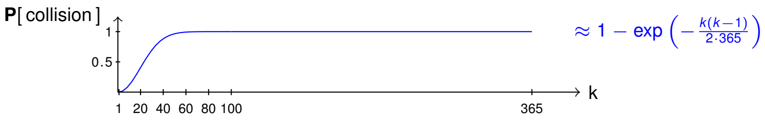
Recall: L_1-distance is

$$\sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right|$$

First Idea might be to approximate each $p_i - \frac{1}{n}$, but this takes at least $\Omega(n)$ queries.

Birthday Paradox:

- If p is (close to) uniform, expect to see collisions after $\approx \sqrt{n}$ samples
- If p is far from uniform, expect to see collisions with ??



High Level Idea

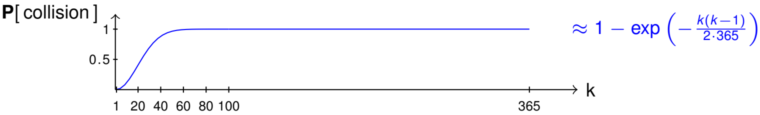
Recall: L_1-distance is

$$\sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right|$$

First Idea might be to approximate each $p_i - \frac{1}{n}$, but this takes at least $\Omega(n)$ queries.

Birthday Paradox:

- If p is (close to) uniform, expect to see collisions after $\approx \sqrt{n}$ samples
- If p is far from uniform, expect to see collisions with even less samples



Collision Probability and L_2-distance

$\|p - u\|_2^2$

Collision Probability and L_2-distance

$$\|p - u\|_2^2 = \sum_{i=1}^{n} (p_i - 1/n)^2$$

Collision Probability and L_2-distance

$$\|p-u\|_2^2 = \sum_{i=1}^{n}(p_i - 1/n)^2 = \sum_{i=1}^{n} p_i^2 - 2\cdot\sum_{i=1}^{n} p_i \cdot \frac{1}{n} + \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2$$

Collision Probability and L_2-distance

$$\|p-u\|_2^2 = \sum_{i=1}^{n}(p_i - 1/n)^2 = \sum_{i=1}^{n} p_i^2 - 2 \cdot \sum_{i=1}^{n} p_i \cdot \frac{1}{n} + \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 = \|p\|_2^2 - \frac{1}{n}$$

Collision Probability and L_2-distance

$$\|p-u\|_2^2 = \sum_{i=1}^{n}(p_i - 1/n)^2 = \sum_{i=1}^{n} p_i^2 - 2\cdot\sum_{i=1}^{n} p_i \cdot \frac{1}{n} + \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 = \|p\|_2^2 - \frac{1}{n}$$

Hence $\|p\|_2^2 = \sum_{i=1}^{n} p_i^2$ captures the L_2-distance to the uniform distribution

Collision Probability and L_2-distance

$$\|p-u\|_2^2 = \sum_{i=1}^{n}(p_i - 1/n)^2 = \sum_{i=1}^{n} p_i^2 - 2\cdot\sum_{i=1}^{n} p_i \cdot \frac{1}{n} + \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 = \|p\|_2^2 - \frac{1}{n}$$

Hence $\|p\|_2^2 = \sum_{i=1}^{n} p_i^2$ captures the L_2-distance to the uniform distribution

APPROXIMATE $\|p\|_2^2$

1. Sample r elements from p, $x_1, x_2, \ldots, x_r \in \{1, \ldots, n\}$
2. For each $1 \leq i < j \leq r$,

$$\sigma_{i,j} := \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases}$$

3. Output $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$.

Collision Probability and L_2-distance

$$\|p-u\|_2^2 = \sum_{i=1}^{n}(p_i - 1/n)^2 = \sum_{i=1}^{n} p_i^2 - 2\cdot\sum_{i=1}^{n} p_i \cdot \frac{1}{n} + \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 = \|p\|_2^2 - \frac{1}{n}$$

Hence $\|p\|_2^2 = \sum_{i=1}^{n} p_i^2$ captures the L_2-distance to the uniform distribution

APPROXIMATE $\|p\|_2^2$

number of samples r will be specified later!

1. Sample r elements from p, $x_1, x_2, \ldots, x_r \in \{1,\ldots,n\}$
2. For each $1 \leq i < j \leq r$,

$$\sigma_{i,j} := \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases}$$

3. Output $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1\leq i<j\leq r} \sigma_{i,j}$.

Runtime Analysis

- Sampling/Query Complexity is obviously r

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:

    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

    $$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \stackrel{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:

    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

    $$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \stackrel{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Proof of $(\star)$:

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j}$$

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:
    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \overset{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Proof of $(\star)$:

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} = \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j}$$

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:
    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \stackrel{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Proof of $(\star)$:

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} = \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j}$$

$$= \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{1}_{x_i = x_j = k}$$

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:
    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \stackrel{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Proof of $(\star)$:

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} = \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j}$$

$$= \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{1}_{x_i = x_j = k} = \sum_{k=1}^{n} \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j = k}$$

Runtime Analysis

- Sampling/Query Complexity is obviously r
- Time Complexity??
  - Evaluating $\sum_{1 \leq i < j \leq r} \sigma_{i,j}$ directly takes time quadratic in r
  - Linear-Time Solution:

    1. Maintain array $A = (a_1, a_2, \ldots, a_n)$, where $a_i \in [0, r]$ counts the frequency of samples of item i
    2. Use formula

    $$\sum_{1 \leq i < j \leq r} \sigma_{i,j} \stackrel{(\star)}{=} \sum_{k=1}^{n} \binom{a_k}{2}$$

    3. Since at most $O(r)$ elements in A will be non-zero, using hash-function allows computation in time $O(r)$

Proof of $(\star)$:

$$\sum_{1 \leq i < j \leq r} \sigma_{i,j} = \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j}$$

$$= \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{1}_{x_i = x_j = k} = \sum_{k=1}^{n} \sum_{1 \leq i < j \leq r} \mathbf{1}_{x_i = x_j = k} = \sum_{k=1}^{n} \binom{a_k}{2}. \quad \square$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\mathbf{E}[\,Y\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\mathbf{E}[\,Y\,] = \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}\,\right]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\begin{aligned}\mathbf{E}[\,Y\,] &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] \\ &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{P}[\,x_i = k\,] \cdot \mathbf{P}[\,x_j = k\,]\end{aligned}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\begin{aligned}
\mathbf{E}[\,Y\,] &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}\,\right] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{P}[\,x_i = k\,] \cdot \mathbf{P}\left[\,x_j = k\,\right] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} p_k^2 = \|p\|_2^2.
\end{aligned}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\begin{aligned}
\mathbf{E}[\,Y\,] &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{P}[\,x_i = k\,] \cdot \mathbf{P}[\,x_j = k\,] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} p_k^2 = \|p\|_2^2.
\end{aligned}$$

- Analysis of the deviation more complex (see next slides):

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\begin{aligned}
\mathbf{E}[\,Y\,] &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}\,\right] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{P}[\,x_i = k\,] \cdot \mathbf{P}\left[\,x_j = k\,\right] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} p_k^2 = \|p\|_2^2.
\end{aligned}$$

- Analysis of the deviation more complex (see next slides):
  - requires a careful analysis of the variance
    (note that the $\sigma_{i,j}$'s are not even pairwise independent! - **Exercise**)

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (1/5):

- Let us start by computing $\mathbf{E}[\,Y\,]$:

$$\begin{aligned}
\mathbf{E}[\,Y\,] &= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} \mathbf{P}[\,x_i = k\,] \cdot \mathbf{P}[\,x_j = k\,] \\
&= \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sum_{k=1}^{n} p_k^2 = \|p\|_2^2.
\end{aligned}$$

- Analysis of the deviation more complex (see next slides):
  - requires a careful analysis of the variance (note that the $\sigma_{i,j}$'s are not even pairwise independent! - **Exercise**)
  - final step is an application of Chebysheff's inequality

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (2/5):

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (2/5):

- Define $\widehat{\sigma}_{i,j} := \sigma_{i,j} - \mathbf{E}[\,\sigma_{i,j}\,]$. Note $\mathbf{E}[\,\widehat{\sigma}_{i,j}\,] = 0$, $\widehat{\sigma}_{i,j} \leq \sigma_{i,j}$ and

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (2/5):

- Define $\widehat{\sigma}_{i,j} := \sigma_{i,j} - \mathbf{E}[\sigma_{i,j}]$. Note $\mathbf{E}[\widehat{\sigma}_{i,j}] = 0$, $\widehat{\sigma}_{i,j} \leq \sigma_{i,j}$ and

$$\mathbf{Var}\left[\sum_{1 \leq i < j \leq r} \sigma_{i,j} \right] = \mathbf{E}\left[\left(\sum_{1 \leq i < j \leq r} \sigma_{i,j} - \sum_{1 \leq i < j \leq r} \mathbf{E}[\sigma_{i,j}] \right)^2 \right] = \mathbf{E}\left[\left(\sum_{1 \leq i < j \leq r} \widehat{\sigma}_{i,j} \right)^2 \right].$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (2/5):

- Define $\widehat{\sigma}_{i,j} := \sigma_{i,j} - \mathbf{E}[\sigma_{i,j}]$. Note $\mathbf{E}[\widehat{\sigma}_{i,j}] = 0$, $\widehat{\sigma}_{i,j} \leq \sigma_{i,j}$ and

$$\mathbf{Var}\left[\sum_{1\leq i<j\leq r} \sigma_{i,j}\right] = \mathbf{E}\left[\left(\sum_{1\leq i<j\leq r} \sigma_{i,j} - \sum_{1\leq i<j\leq r} \mathbf{E}[\sigma_{i,j}]\right)^2\right] = \mathbf{E}\left[\left(\sum_{1\leq i<j\leq r} \widehat{\sigma}_{i,j}\right)^2\right].$$

- Expanding yields:

There are $2 \cdot \binom{3}{2} = 6$ combinations!

$$\underbrace{\sum_{1\leq i<j\leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2\right]}_{=A} + \underbrace{\sum_{i,j,k,\ell \text{ diff.}} \mathbf{E}[\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}]}_{=B} + 6 \cdot \underbrace{\sum_{1\leq i<j<k\leq r} \mathbf{E}[\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{j,k}]}_{=C}.$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,] = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j}\,] \cdot \mathbf{E}[\,\widehat{\sigma}_{k,\ell}\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\sigma_{i,j}^2\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,] = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j}\,] \cdot \mathbf{E}[\,\widehat{\sigma}_{k,\ell}\,] = 0.$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2 \right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\sigma_{i,j}^2 \right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\, \sigma_{i,j} \,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\, \widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell} \,] = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\, \widehat{\sigma}_{i,j} \,] \cdot \mathbf{E}[\, \widehat{\sigma}_{k,\ell} \,] = 0.$$

$$C = \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\, \widehat{\sigma}_{i,j} \widehat{\sigma}_{i,k} \,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\widehat{\sigma}_{i,j}^2\,\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\,\sigma_{i,j}^2\,\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,j,k,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,] = \sum_{i,j,k,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j}\,] \cdot \mathbf{E}[\,\widehat{\sigma}_{k,\ell}\,] = 0.$$

Covariance Formula: $\mathbf{E}[\,(X - \mathbf{E}[\,X\,])(Y - \mathbf{E}[\,Y\,])\,] = \mathbf{E}[\,XY\,] - \mathbf{E}[\,X\,]\,\mathbf{E}[\,Y\,]$

$$C = \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\,\widehat{\sigma}_{i,j}\widehat{\sigma}_{i,k}\,] \leq \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\,\sigma_{i,j}\sigma_{i,k}\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2\right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\sigma_{i,j}^2\right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,j,k,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,] = \sum_{i,j,k,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j}\,] \cdot \mathbf{E}[\,\widehat{\sigma}_{k,\ell}\,] = 0.$$

Covariance Formula: $\mathbf{E}[\,(X - \mathbf{E}[\,X\,])(Y - \mathbf{E}[\,Y\,])\,] = \mathbf{E}[\,XY\,] - \mathbf{E}[\,X\,]\,\mathbf{E}[\,Y\,]$

$$C = \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\,\widehat{\sigma}_{i,j}\widehat{\sigma}_{i,k}\,] \leq \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\,\sigma_{i,j}\sigma_{i,k}\,]$$

$$= \sum_{1 \leq i < j < k \leq r} \sum_{\ell \in [n]} \mathbf{P}[\,X_i = X_j = X_k = \ell\,]$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2 \right] \leq \sum_{1 \leq i < j \leq r} \mathbf{E}\left[\sigma_{i,j}^2 \right] = \sum_{1 \leq i < j \leq r} \mathbf{E}[\, \sigma_{i,j} \,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\, \widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell} \,] = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\, \widehat{\sigma}_{i,j} \,] \cdot \mathbf{E}[\, \widehat{\sigma}_{k,\ell} \,] = 0.$$

Covariance Formula: $\mathbf{E}[\,(X - \mathbf{E}[\,X\,])(Y - \mathbf{E}[\,Y\,])\,] = \mathbf{E}[\,XY\,] - \mathbf{E}[\,X\,]\,\mathbf{E}[\,Y\,]$

$$C = \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\, \widehat{\sigma}_{i,j} \widehat{\sigma}_{i,k} \,] \leq \sum_{1 \leq i < j < k \leq r} \mathbf{E}[\, \sigma_{i,j} \sigma_{i,k} \,]$$

$$= \sum_{1 \leq i < j < k \leq r} \sum_{\ell \in [n]} \mathbf{P}[\, X_i = X_j = X_k = \ell \,] = \binom{r}{3} \cdot \sum_{\ell \in [n]} p_\ell^3$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (3/5):

$$A = \sum_{1\leq i<j\leq r} \mathbf{E}\left[\widehat{\sigma}_{i,j}^2\right] \leq \sum_{1\leq i<j\leq r} \mathbf{E}\left[\sigma_{i,j}^2\right] = \sum_{1\leq i<j\leq r} \mathbf{E}[\,\sigma_{i,j}\,] = \binom{r}{2} \cdot \|p\|_2^2.$$

$$B = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j} \cdot \widehat{\sigma}_{k,\ell}\,] = \sum_{i,\,j,\,k,\,\ell \text{ diff.}} \mathbf{E}[\,\widehat{\sigma}_{i,j}\,] \cdot \mathbf{E}[\,\widehat{\sigma}_{k,\ell}\,] = 0.$$

Covariance Formula: $\mathbf{E}[\,(X - \mathbf{E}[\,X\,])(Y - \mathbf{E}[\,Y\,])\,] = \mathbf{E}[\,XY\,] - \mathbf{E}[\,X\,]\,\mathbf{E}[\,Y\,]$

$$C = \sum_{1\leq i<j<k\leq r} \mathbf{E}[\,\widehat{\sigma}_{i,j}\widehat{\sigma}_{i,k}\,] \leq \sum_{1\leq i<j<k\leq r} \mathbf{E}[\,\sigma_{i,j}\sigma_{i,k}\,]$$

$$= \sum_{1\leq i<j<k\leq r} \sum_{\ell\in[n]} \mathbf{P}[\,X_i = X_j = X_k = \ell\,] = \binom{r}{3} \cdot \sum_{\ell\in[n]} p_\ell^3 \leq \frac{\sqrt{3}}{2} \left(\binom{r}{2} \|p\|_2^2\right)^{3/2}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (4/5):

- We have just shown that:

$$\mathbf{Var}\left[\sum_{1 \leq i < j \leq r} \sigma_{i,j}\right] = A + B + 6C$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (4/5):

- We have just shown that:

$$\mathbf{Var}\left[\sum_{1 \leq i < j \leq r} \sigma_{i,j}\right] = A + B + 6C$$

$$= \binom{r}{2} \cdot \|p\|_2^2 + 0 + 6 \cdot \frac{\sqrt{3}}{2}\left(\binom{r}{2}\|p\|_2^2\right)^{3/2}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left| Y - \|p\|_2^2 \right| \geq \epsilon \cdot \|p\|_2^2 \right] \leq 1/3.$$

Proof (4/5):

- We have just shown that:

$$\begin{aligned}
\mathbf{Var}\left[\sum_{1 \leq i < j \leq r} \sigma_{i,j} \right] &= A + B + 6C \\
&= \binom{r}{2} \cdot \|p\|_2^2 + 0 + 6 \cdot \frac{\sqrt{3}}{2} \left(\binom{r}{2} \|p\|_2^2 \right)^{3/2} \\
&\leq 5 \left(\binom{r}{2} \|p\|_2^2 \right)^{3/2}
\end{aligned}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\right] \leq 1/3.$$

Proof (5/5):

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (5/5):

- Applying Chebyshef's inequality to $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$ yields:

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (5/5):

- Applying Chebyshef's inequality to $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$ yields:

$$\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq \frac{\mathbf{Var}\,[\,Y\,]}{\epsilon^2 \cdot \|p\|_2^4}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (5/5):

- Applying Chebyshef's inequality to $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$ yields:

$$\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq \frac{\mathbf{Var}\,[\,Y\,]}{\epsilon^2 \cdot \|p\|_2^4}$$

$$\leq \frac{\frac{1}{\binom{r}{2}^2} \cdot 7\left(\binom{r}{2} \cdot \|p\|_2^2\right)^{3/2}}{\epsilon^2 \cdot \|p\|_2^4}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (5/5):

- Applying Chebyshef's inequality to $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$ yields:

$$\begin{aligned}
\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]| \geq \epsilon \cdot \|p\|_2^2\,\right] &\leq \frac{\mathbf{Var}[\,Y\,]}{\epsilon^2 \cdot \|p\|_2^4} \\
&\leq \frac{\frac{1}{\binom{r}{2}^2} \cdot 7\left(\binom{r}{2} \cdot \|p\|_2^2\right)^{3/2}}{\epsilon^2 \cdot \|p\|_2^4} \\
&\leq \frac{14}{r \cdot \|p\|_2 \cdot \epsilon^2}
\end{aligned}$$

Approximation Analysis

Analysis

For any value $r \geq 42 \cdot \frac{\sqrt{n}}{\epsilon^2}$, the algorithm returns a value Y such that

$$\mathbf{P}\left[\,\left|Y - \|p\|_2^2\right| \geq \epsilon \cdot \|p\|_2^2\,\right] \leq 1/3.$$

Proof (5/5):

- Applying Chebyshef's inequality to $Y := \frac{1}{\binom{r}{2}} \cdot \sum_{1 \leq i < j \leq r} \sigma_{i,j}$ yields:

$$\begin{aligned}
\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]| \geq \epsilon \cdot \|p\|_2^2\,\right] &\leq \frac{\mathbf{Var}[\,Y\,]}{\epsilon^2 \cdot \|p\|_2^4} \\
&\leq \frac{\frac{1}{\binom{r}{2}^2} \cdot 7\left(\binom{r}{2} \cdot \|p\|_2^2\right)^{3/2}}{\epsilon^2 \cdot \|p\|_2^4} \\
&\leq \frac{14}{r \cdot \|p\|_2 \cdot \epsilon^2} \\
&\leq \frac{14}{r \cdot (1/\sqrt{n}) \cdot \epsilon^2} \qquad \square
\end{aligned}$$

Approximation of $\|p - u\|_1$ using $\|p\|_2^2$

UNIFORM-TEST

1. Run **APPROXIMATE** $\|p\|_2^2$ with $r = 42 \cdot \frac{\sqrt{n}}{(\epsilon^2/4)^2} = \mathcal{O}(\frac{\sqrt{n}}{\epsilon^4})$ samples to get a value Y such that

$$\mathbf{P}\left[\, |Y - \mathbf{E}[\,Y\,]\,| \geq \epsilon^2/4 \cdot \|p\|_2^2 \,\right] \leq 1/3.$$

2. If $Y \geq \frac{1+\epsilon^2/2}{n}$, then REJECT.
3. Otherwise, ACCEPT.

Approximation of $\|p-u\|_1$ using $\|p\|_2^2$

UNIFORM-TEST

1. Run **APPROXIMATE** $\|p\|_2^2$ with $r = 42 \cdot \frac{\sqrt{n}}{(\epsilon^2/4)^2} = \mathcal{O}(\frac{\sqrt{n}}{\epsilon^4})$ samples to get a value Y such that

$$\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]\,| \geq \epsilon^2/4 \cdot \|p\|_2^2\,\right] \leq 1/3.$$

2. If $Y \geq \frac{1+\epsilon^2/2}{n}$, then REJECT.
3. Otherwise, ACCEPT.

Correctness Analysis

- If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq 2/3$.
- If p is ϵ-far from u, i.e., $\sum_{i=1}^{n} \left|p_i - \frac{1}{n}\right| \geq \epsilon$), then $\mathbf{P}[\,\text{REJECT}\,] \geq 2/3$.

Approximation of $\|p - u\|_1$ using $\|p\|_2^2$

UNIFORM-TEST

1. Run **APPROXIMATE** $\|p\|_2^2$ with $r = 42 \cdot \frac{\sqrt{n}}{(\epsilon^2/4)^2} = \mathcal{O}(\frac{\sqrt{n}}{\epsilon^4})$ samples to get a value Y such that
$$\mathbf{P}\left[\,|Y - \mathbf{E}[\,Y\,]\,| \geq \epsilon^2/4 \cdot \|p\|_2^2\,\right] \leq 1/3.$$
2. If $Y \geq \frac{1+\epsilon^2/2}{n}$, then REJECT.
3. Otherwise, ACCEPT.

Correctness Analysis

- If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq 2/3$.
- If p is ϵ-far from u, i.e., $\sum_{i=1}^{n} \left|p_i - \frac{1}{n}\right| \geq \epsilon$), then $\mathbf{P}[\,\text{REJECT}\,] \geq 2/3$.

Exercise: Prove that any testing algorithm in this model will have a two-sided error!

Analysis of UNIFORM-TEST (1/2)

Case 1: p is uniform.

In this case

$$\|p\|_2^2 = \frac{1}{n},$$

Analysis of UNIFORM-TEST (1/2)

Case 1: p is uniform.

In this case

$$\|p\|_2^2 = \frac{1}{n},$$

and the approximation guarantee on Y implies

$$\mathbf{P}\Big[\, Y \geq \|p\|_2^2 \cdot (1 + \epsilon^2/4) \,\Big] \leq 1/3,$$

Analysis of UNIFORM-TEST (1/2)

Case 1: p is uniform.

In this case

$$\|p\|_2^2 = \frac{1}{n},$$

and the approximation guarantee on Y implies

$$\mathbf{P}\left[\, Y \geq \|p\|_2^2 \cdot (1 + \epsilon^2/4) \,\right] \leq 1/3,$$

which means that the algorithm will ACCEPT with probability at least $2/3$.

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\text{ REJECT }] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\text{ REJECT }] \leq 2/3$ implies

$$\mathbf{P}\left[Y > \frac{1+\epsilon^2/2}{n} \right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n} \cdot (1+\epsilon^2/2) \cdot \frac{1}{1-\epsilon^2/4}$$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n} \cdot (1+\epsilon^2/2) \cdot \frac{1}{1-\epsilon^2/4}$$

$1 \leq (1+\epsilon^2/3) \cdot (1-\epsilon^2/4)$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n} \cdot (1+\epsilon^2/2) \cdot \frac{1}{1-\epsilon^2/4} \leq \frac{1+\epsilon^2}{n}.$$

$1 \leq (1+\epsilon^2/3) \cdot (1-\epsilon^2/4)$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4)\cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n}\cdot(1+\epsilon^2/2)\cdot\frac{1}{1-\epsilon^2/4} \leq \frac{1+\epsilon^2}{n}.$$

$1 \leq (1+\epsilon^2/3)\cdot(1-\epsilon^2/4)$

Hence,

$$\|p-u\|_2^2 = \|p\|_2^2 - \frac{1}{n} < \frac{\epsilon^2}{n} \qquad \Rightarrow \qquad \|p-u\|_2 < \frac{\epsilon}{\sqrt{n}}.$$

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\,Y > \frac{1+\epsilon^2/2}{n}\,\right] < 2/3. \qquad (1)$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\,Y > (1-\epsilon^2/4)\cdot \|p\|_2^2\,\right] \geq 2/3. \qquad (2)$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n}\cdot(1+\epsilon^2/2)\cdot\frac{1}{1-\epsilon^2/4} \leq \frac{1+\epsilon^2}{n}.$$

$1 \leq (1+\epsilon^2/3)\cdot(1-\epsilon^2/4)$

Hence,

$$\|p-u\|_2^2 = \|p\|_2^2 - \frac{1}{n} < \frac{\epsilon^2}{n} \qquad \Rightarrow \qquad \|p-u\|_2 < \frac{\epsilon}{\sqrt{n}}.$$

Since $\|.\|_2 \geq \frac{1}{\sqrt{n}}\cdot\|.\|_1$,

Analysis of UNIFORM-TEST (2/2)

Case 2: p is ϵ-far from u.

We will show that if $\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$, then p is ϵ-close to u.

$\mathbf{P}[\,\text{REJECT}\,] \leq 2/3$ implies

$$\mathbf{P}\left[\, Y > \frac{1+\epsilon^2/2}{n} \,\right] < 2/3. \tag{1}$$

From line 1 of the algorithm we know that

$$\mathbf{P}\left[\, Y > (1-\epsilon^2/4) \cdot \|p\|_2^2 \,\right] \geq 2/3. \tag{2}$$

Combining (1) and (2) yields, and rearranging yields

$$\|p\|_2^2 < \frac{1}{n} \cdot (1+\epsilon^2/2) \cdot \frac{1}{1-\epsilon^2/4} \leq \frac{1+\epsilon^2}{n}.$$

$1 \leq (1+\epsilon^2/3) \cdot (1-\epsilon^2/4)$

Hence,

$$\|p-u\|_2^2 = \|p\|_2^2 - \frac{1}{n} < \frac{\epsilon^2}{n} \qquad \Rightarrow \qquad \|p-u\|_2 < \frac{\epsilon}{\sqrt{n}}.$$

Since $\|.\|_2 \geq \frac{1}{\sqrt{n}} \cdot \|.\|_1$,

$$\|p-u\|_1 \leq \sqrt{n} \cdot \|p-u\|_2 < \epsilon. \qquad \square$$

Outline

Introduction

Upper Bounds on Testing Uniformity

Lower Bounds on Testing Uniformity

Extensions

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1$

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1 = \epsilon \cdot n \cdot 1/n = \epsilon$.

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1 = \epsilon \cdot n \cdot 1/n = \epsilon$.
- E.g., $n = 16$, $\epsilon = 1/4$, $\mathcal{I} = \{1, 4\}$:

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1 = \epsilon \cdot n \cdot 1/n = \epsilon$.
- E.g., $n = 16$, $\epsilon = 1/4$, $\mathcal{I} = \{1, 4\}$:

$$p = \Bigg(\underbrace{\frac{2}{n}, 0, 0, \frac{2}{n}}_{\epsilon n = 4 \text{ elements}}, \underbrace{\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}}_{12 \text{ elements}} \Bigg)$$

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1 = \epsilon \cdot n \cdot 1/n = \epsilon$.
- E.g., $n = 16$, $\epsilon = 1/4$, $\mathcal{I} = \{1, 4\}$:

Idea is that algorithm needs enough samples of the first $\epsilon \cdot n$ elements to see any collisions!

$$p = \Bigg(\underbrace{\frac{2}{n}, 0, 0, \frac{2}{n}}_{\epsilon n = 4 \text{ elements}}, \underbrace{\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}}_{12 \text{ elements}} \Bigg)$$

Lower Bound

Theorem

Let $0 < \epsilon < 1$. There is no algorithm with the following three properties:

1. The algorithm samples at most $r := \frac{1}{64}\sqrt{n/\epsilon}$ times from p,
2. If $p = u$, then $\mathbf{P}[\,\text{ACCEPT}\,] \geq \frac{2}{3}$,
3. If $\|p - u\|_1 \geq \epsilon$, then $\mathbf{P}[\,\text{REJECT}\,] \geq \frac{2}{3}$.

Exercise: Can you see why is it important to choose $\mathcal{I}$ **randomly**?

Proof Outline.

- Generate a distribution p randomly as follows:
  - Pick a set $\mathcal{I} \subseteq \{1, \ldots, \epsilon \cdot n\}$ of size $\epsilon \cdot n/2$ uniformly at random.
  - Then define:

$$p_i = \begin{cases} \frac{2}{n} & \text{if } i \in \mathcal{I}, \\ 0 & \text{if } i \in \{1, \ldots, \epsilon \cdot n\} \setminus \mathcal{I}, \\ \frac{1}{n} & \text{if } \epsilon \cdot n < i < n. \end{cases}$$

  - Then $\|p - u\|_1 = \epsilon \cdot n \cdot 1/n = \epsilon$.
- E.g., $n = 16$, $\epsilon = 1/4$, $\mathcal{I} = \{1, 4\}$:

Idea is that algorithm needs enough samples of the first $\epsilon \cdot n$ elements to see any collisions!

$$p = \Bigg(\underbrace{\frac{2}{n}, 0, 0, \frac{2}{n}}_{\epsilon n = 4 \text{ elements}}, \underbrace{\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}}_{12 \text{ elements}} \Bigg)$$

Exercise (in Class)

What is the **expected number of collisions** among $\mathcal{I}$, if we are sampling t times from the distribution p as described on the previous slide?

Let Z denote the number of collisions, so $Z = \sum_{1 \leq k < \ell \leq t} \mathbf{1}_{x_k = x_\ell \wedge x_k \in \mathcal{I}}$. Taking expectations and using linearity of expectation yields:

$$
\begin{aligned}
\mathbf{E}[\,Z\,] = \mathbf{E}\left[\sum_{1 \leq k < \ell \leq t} \mathbf{1}_{x_k = x_\ell \wedge x_k \in \mathcal{I}}\right] &= \sum_{1 \leq k < \ell \leq t} \mathbf{E}\left[\,\mathbf{1}_{x_k = x_\ell \wedge x_k \in \mathcal{I}}\,\right] \\
&= \sum_{1 \leq k < \ell \leq t} \mathbf{P}[\,x_k = x_\ell \wedge x_k \in \mathcal{I}\,] \cdot 1 \\
&= \sum_{1 \leq k < \ell \leq t} \sum_{i=1}^{\epsilon n} \mathbf{P}[\,x_k = x_\ell \wedge x_k \in \mathcal{I} \wedge x_k = i\,] \\
&= \sum_{1 \leq k < \ell \leq t} \sum_{i=1}^{\epsilon n} \mathbf{P}[\,i \in \mathcal{I}\,] \cdot \mathbf{P}[\,x_k = x_\ell = i \mid i \in \mathcal{I}\,] \\
&= \sum_{1 \leq k < \ell \leq t} \sum_{i=1}^{\epsilon n} \frac{1}{2} \cdot \left(\frac{2}{n}\right)^2 = \binom{t}{2} \cdot \epsilon n \cdot \left(\frac{2}{n}\right)^2.
\end{aligned}
$$

Hence if $t = o(\sqrt{n/\epsilon})$, then $\mathbf{E}[\,Z\,] \to 0$ and thus $\mathbf{P}[\,Z = 0\,] \to 1$, as $n \to \infty$.

Outline

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$\|p - q\|_2^2$

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$$\|p - q\|_2^2 = \sum_{i=1}^{n} (p_i - q_i)^2$$

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$$\|p - q\|_2^2 = \sum_{i=1}^{n} (p_i - q_i)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} q_i^2 - 2 \cdot \sum_{i=1}^{n} p_i \cdot q_i$$

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$$\|p-q\|_2^2 = \sum_{i=1}^{n}(p_i - q_i)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} q_i^2 - 2\cdot\sum_{i=1}^{n} p_i \cdot q_i$$
$$= \|p\|_2^2 + \|q\|_2^2 - 2\cdot\langle p, q\rangle$$

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$$\|p-q\|_2^2 = \sum_{i=1}^{n}(p_i - q_i)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} q_i^2 - 2 \cdot \sum_{i=1}^{n} p_i \cdot q_i$$
$$= \|p\|_2^2 + \|q\|_2^2 - 2 \cdot \langle p, q \rangle$$

We already know how to estimate $\|p\|_2^2$ and $\|q\|_2^2$!

Extension 1: Testing Closeness of Arbitrary Distributions (1/2)

$$\|p-q\|_2^2 = \sum_{i=1}^{n}(p_i-q_i)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} q_i^2 - 2\cdot\sum_{i=1}^{n} p_i\cdot q_i$$
$$= \|p\|_2^2 + \|q\|_2^2 - 2\cdot\langle p,q\rangle$$

We already know how to estimate $\|p\|_2^2$ and $\|q\|_2^2$!

APPROXIMATE $\langle p,q\rangle$

1. Sample r elements from p, $x_1, x_2, \ldots, x_r \in [n]$, and sample r elements from q, $y_1, y_2, \ldots, y_r \in [n]$
2. For each $1 \leq i < j \leq r$,

$$\tau_{i,j} := \begin{cases} 1 & \text{if } x_i = y_j, \\ 0 & \text{otherwise.} \end{cases}$$

3. Output $Y := \frac{1}{r^2}\sum_{1\leq i,j\leq r}\tau_{i,j}$.

Extension 1: Testing Closeness of Arbitrary Distributions (2/2)

Theorem (Batu, Fortnow, Rubinfeld, Smith, White; JACM 60(1), 2013)

There exists an algorithm using $\mathcal{O}(1/\epsilon^4)$ samples such that if the distributions p and q satisfy $\|p - q\|_2 \leq \epsilon/2$, then the algorithm accepts with probability at least $2/3$. If $\|p - q\|_2 \geq \epsilon$, then the algorithm rejects with probability at least $2/3$.

Extension 1: Testing Closeness of Arbitrary Distributions (2/2)

Theorem (Batu, Fortnow, Rubinfeld, Smith, White; JACM 60(1), 2013)

There exists an algorithm using $\mathcal{O}(1/\epsilon^4)$ samples such that if the distributions p and q satisfy $\|p - q\|_2 \leq \epsilon/2$, then the algorithm accepts with probability at least $2/3$. If $\|p - q\|_2 \geq \epsilon$, then the algorithm rejects with probability at least $2/3$.

Theorem (Batu, Fortnow, Rubinfeld, Smith, White; JACM 60(1), 2013)

There exists an algorithm using $\mathcal{O}(1/\epsilon^4 \cdot n^{2/3} \log n)$ samples such that if the distributions p and q satisfy $\|p - q\|_1 \leq \max\{\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}}\}$, then the algorithm accepts with probability at least $2/3$. If $\|p - q\|_1 \geq \epsilon$, then the algorithm rejects with probability at least $2/3$.

Extension 1: Testing Closeness of Arbitrary Distributions (2/2)

Theorem (Batu, Fortnow, Rubinfeld, Smith, White; JACM 60(1), 2013)

There exists an algorithm using $\mathcal{O}(1/\epsilon^4)$ samples such that if the distributions p and q satisfy $\|p-q\|_2 \leq \epsilon/2$, then the algorithm accepts with probability at least $2/3$. If $\|p-q\|_2 \geq \epsilon$, then the algorithm rejects with probability at least $2/3$.

Theorem (Batu, Fortnow, Rubinfeld, Smith, White; JACM 60(1), 2013)

There exists an algorithm using $\mathcal{O}(1/\epsilon^4 \cdot n^{2/3} \log n)$ samples such that if the distributions p and q satisfy $\|p-q\|_1 \leq \max\{\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}}\}$, then the algorithm accepts with probability at least $2/3$. If $\|p-q\|_1 \geq \epsilon$, then the algorithm rejects with probability at least $2/3$.

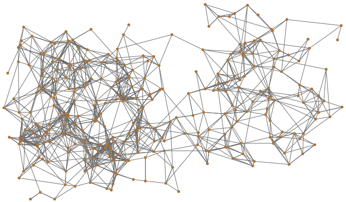
	L_2-distance	L_1-distance
Testing uniformity $\|p-u\|$	$\Theta(1)$	$\Theta(\sqrt{n})$
Testing closeness $\|p-q\|$	$\Theta(1)$	$\in [\Omega(n^{2/3}), \mathcal{O}(n^{2/3} \log n)]$

Figure: Overview of the known sampling complexities for constant $\epsilon \in (0,1)$.

Extension 2: Testing Conductance of Graphs

Testing Conductance of Graphs

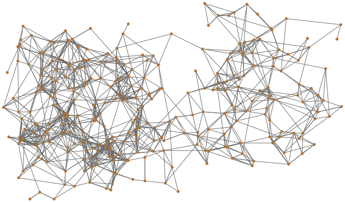
- Idea: Start several random walks from the same vertex



Extension 2: Testing Conductance of Graphs

Testing Conductance of Graphs

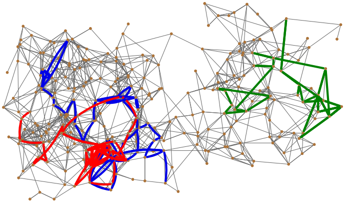
- Idea: Start several random walks from the same vertex
- Count the number of pairwise collisions among the endpoints of the walks



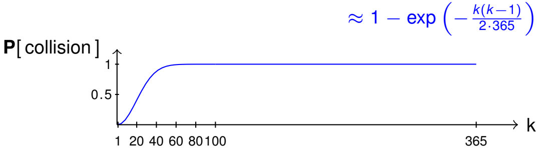
Extension 2: Testing Conductance of Graphs

Testing Conductance of Graphs

- Idea: Start several random walks from the same vertex
- Count the number of pairwise collisions among the endpoints of the walks
  - If the number of collisions high, graphs is not an expander
  - If the number of collisions is sufficiently small, graph is an expander

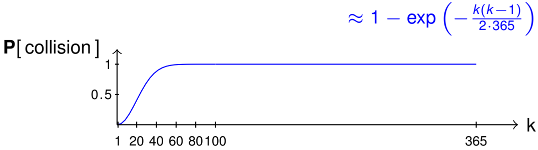


Application 3: Estimating Population Sizes using Mark & Recapture



Source: Wikipedia

Application 3: Estimating Population Sizes using Mark & Recapture

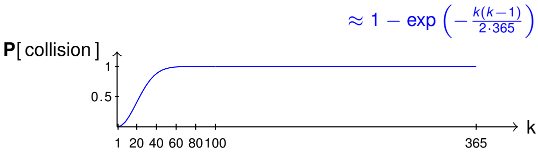


Source: Wikipedia

Mark & Recapture Method:

- First phase: A portion of the population is captured, marked and released
- Second phase: Another portion is captured and the number of marked individuals is counted

Application 3: Estimating Population Sizes using Mark & Recapture



Source: Wikipedia

Mark & Recapture Method:

- First phase: A portion of the population is captured, marked and released
- Second phase: Another portion is captured and the number of marked individuals is counted

  - Essentially the same as collision sampling
  - Can be used to estimate the size of a large network if each node has a unique ID (within an unknown range)