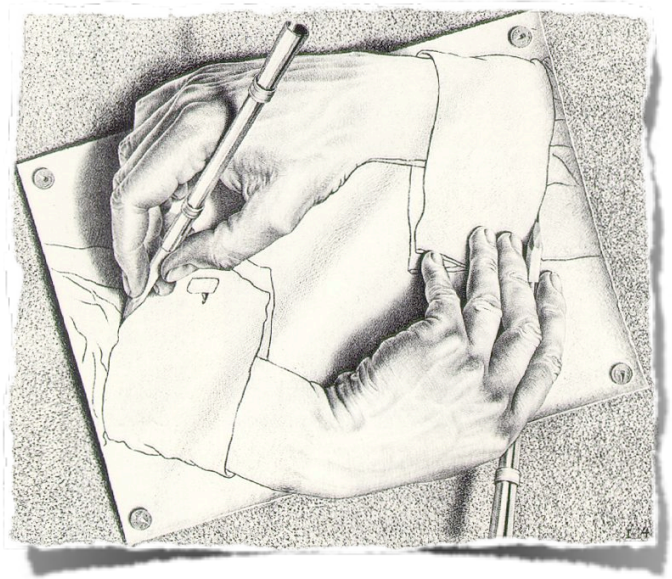
Sketching and Streaming for Distributions

Piotr Indyk
Massachusetts Institute of Technology

Andrew McGregor
University of California, San Diego



Main Material:

Stable distributions, pseudo-random generators, embeddings, and data stream computation
Piotr Indyk (FOCS 2000)

Sketching information divergences
Sudipto Guha, Piotr Indyk, Andrew McGregor (COLT 2007)

Declaring independence via the sketching of sketches
Piotr Indyk, Andrew McGregor (SODA 2008)

The Problem

The Problem

- List of *m* red values and *m* green values in [*n*]

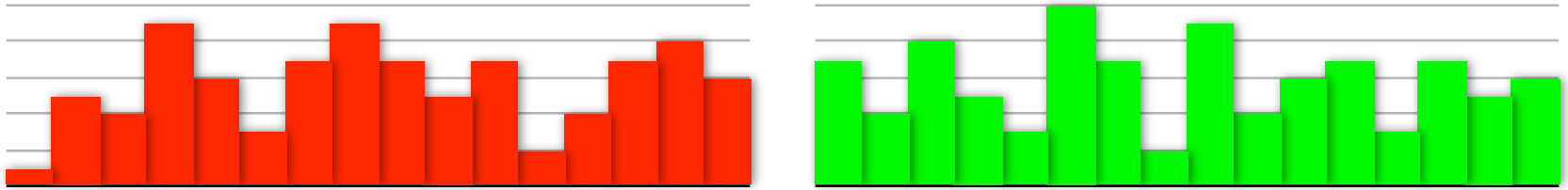
3,5,3,7,5,4,8,5,3,7,5,4,8,6,3,2,6,4,7,3,4, ...

The Problem

- List of m red values and m green values in $[n]$

  3,5,3,7,5,4,8,5,3,7,5,4,8,6,3,2,6,4,7,3,4, ...

- Define distributions $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$

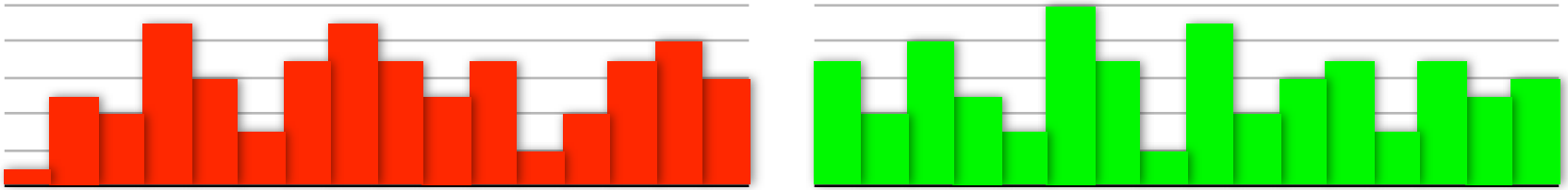


The Problem

- List of m red values and m green values in $[n]$

  3,5,3,7,5,4,8,5,3,7,5,4,8,6,3,2,6,4,7,3,4, ...

- Define distributions $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$



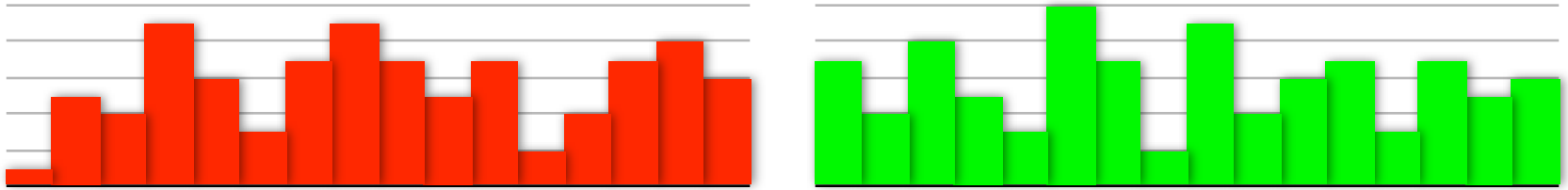
- How "different" are p and q?

The Problem

- List of m red values and m green values in $[n]$

  3,5,3,7,5,4,8,5,3,7,5,4,8,6,3,2,6,4,7,3,4, ...

- Define distributions $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$



- How "different" are p and q?

Variational: $\sum |p_i - q_i|$ Kullback-Leibler: $\sum p_i \log(p_i/q_i)$

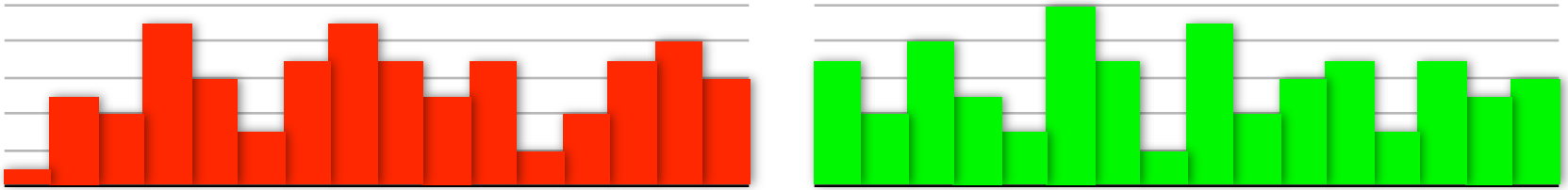
Euclidean: $\sum (p_i - q_i)^2$ Hellinger: $\sum (\sqrt{p_i} - \sqrt{q_i})^2$

The Problem

- List of m red values and m green values in $[n]$

  3,5,3,7,5,4,8,5,3,7,5,4,8,6,3,2,6,4,7,3,4, ...

- Define distributions $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$



- How "different" are p and q?

$$D_f(p, q) = \sum p_i f(q_i/p_i)$$

$$B_F(p, q) = \sum [F(p_i) - F(q_i) - (p_i - q_i)F'(q_i)]$$

where f and F are convex and $f(1)=0$.

The Catch...

The Catch...

- What if *m* and *n* are huge and you can't store the list?

The Catch...

- What if *m* and *n* are huge and you can't store the list?
- *Applications:* monitoring internet traffic, I/O efficient external memory, processing huge log files, database query planning, sensor networks, ...

The Catch...

- What if *m* and *n* are huge and you can't store the list?
- *Applications:* monitoring internet traffic, I/O efficient external memory, processing huge log files, database query planning, sensor networks, ...
- *Data Stream Model:*

  No control over the order of the stream

  Limited working memory, e.g., polylog(n,m) space

  Limited time to process each element

The Catch...

- What if *m* and *n* are huge and you can't store the list?
- *Applications:* monitoring internet traffic, I/O efficient external memory, processing huge log files, database query planning, sensor networks, ...
- *Data Stream Model:*

  No control over the order of the stream

  Limited working memory, e.g., polylog(n,m) space

  Limited time to process each element
- *Previous work:* quantiles, frequency moments, histograms, clustering, entropy, graph problems...

  see, e.g., Muthukrishnan "Data Streams: Algorithms and Applications"

Today's Talk

Today's Talk

- *Sketching L_p distances ($0<p\leq 2$):*

  $(1+\epsilon)$-approx. with prob. $1-\delta$ in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space

  Stable distributions and pseudo-random generators

  Stable distributions, pseudo-random generators, embeddings & data stream computation (Indyk, FOCS 2000)

Today's Talk

- *Sketching L_p distances ($0<p\leq 2$):*

  $(1+\epsilon)$-approx. with prob. $1-\delta$ in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space

  Stable distributions and pseudo-random generators

  Stable distributions, pseudo-random generators, embeddings & data stream computation (Indyk, FOCS 2000)

- *Impossibility of Extending to Other Divergences:*

  Can we sketch other divergences such as Hellinger?

  Lower bounds via communication complexity

  Sketching information divergences (Guha, Indyk, McGregor, COLT 2007)

Today's Talk

- *Sketching L_p distances ($0<p\leq 2$):*

  $(1+\epsilon)$-approx. with prob. $1-\delta$ in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space

  Stable distributions and pseudo-random generators

  Stable distributions, pseudo-random generators, embeddings & data stream computation (Indyk, FOCS 2000)

- *Impossibility of Extending to Other Divergences:*

  Can we sketch other divergences such as Hellinger?

  Lower bounds via communication complexity

  Sketching information divergences (Guha, Indyk, McGregor, COLT 2007)

- *Using sketches to test independence:*

  Testing independence between data streams

  Declaring independence via the sketching of sketches (Indyk, McGregor, SODA 2008)

1. Sketching L_p distances

p-stable distributions, pseudo-random generators

2. The Unsketchables

information divergences, communication complexity

3. Sketching Sketches

identifying correlations in data streams

1. Sketching L_p distances

p-stable distributions, pseudo-random generators

2. The Unsketchables

information divergences, communication complexity

3. Sketching Sketches

identifying correlations in data streams

Stable Distributions

Stable Distributions

- A *p-stable distribution* μ has the following property:

If $X, Y, Z \sim \mu$ and $a, b \in \mathbb{R}$ then :

$$aX + bY \sim (|a|^p + |b|^p)^{1/p} Z$$

Stable Distributions

- A *p-stable distribution* μ has the following property:

$$\text{If } X, Y, Z \sim \mu \text{ and } a, b \in \mathbb{R} \text{ then :}$$
$$aX + bY \sim (|a|^p + |b|^p)^{1/p} Z$$

- *Examples:*

Normal(0,1) is 2-stable: $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Cauchy is 1-stable: $\frac{1}{\pi} \frac{1}{1+x^2}$

Approximating L_1 and L_2

Approximating L_1 and L_2

- Let μ be a p-stable distribution ($0<p\leq 1$)

Approximating L_1 and L_2

- Let μ be a p-stable distribution ($0<p\leq 1$)
- *Ideal Algorithm:*

```
For i = 1 to k:
    Let x be a length n vector with x_j ~ μ
    Compute t_i = |x.(p-q)|
Return median(t_1, t_2, ... , t_n)/median(|μ|)
```

Approximating L_1 and L_2

- Let μ be a p-stable distribution ($0<p\leq 1$)
- *Ideal Algorithm:*

```
For i = 1 to k:
    Let x be a length n vector with x_j ~ μ
    Compute t_i = |x.(p-q)|
Return median(t_1, t_2, ... , t_n)/median(|μ|)
```

Easy to compute $x.(p\text{-}q)$: for stream 3,5,3,7,5, ... compute $x_3\text{-}x_5+x_3\text{-}x_7\text{-}x_5$- ... and scale.

Approximating L_1 and L_2

- Let μ be a p-stable distribution ($0<p\leq 1$)
- *Ideal Algorithm:*

```
For i = 1 to k:
    Let x be a length n vector with x_j ~ μ
    Compute t_i = |x.(p-q)|
Return median(t_1, t_2, ... , t_n)/median(|μ|)
```

  Easy to compute $x.(p\text{-}q)$: for stream 3,5,3,7,5, ... compute $x_3\text{-}x_5+x_3\text{-}x_7\text{-}x_5$- ... and scale.

- *Lemma:* Returns $(1\pm\epsilon)L_p(p\text{-}q)$ with prob. $1\text{-}\delta$, if $k=\tilde{O}(\epsilon^{-2}\ln\delta^{-1})$.

Approximating L_1 and L_2

- Let μ be a p-stable distribution ($0<p\leq 1$)
- *Ideal Algorithm:*

```
For i = 1 to k:
    Let x be a length n vector with x_j ~ μ
    Compute t_i = |x.(p-q)|
Return median(t_1, t_2, ... , t_n)/median(|μ|)
```

  Easy to compute $x.(p\text{-}q)$: for stream 3,5,3,7,5, ... compute $x_3\text{-}x_5+x_3\text{-}x_7\text{-}x_5$- ... and scale.

- *Lemma:* Returns $(1\pm\epsilon)L_p(p\text{-}q)$ with prob. $1\text{-}\delta$, if $k=\tilde{O}(\epsilon^{-2}\ln\delta^{-1})$.
- *Proof:*

  Each $t_i \sim L_1(p\text{-}q)\,|\mu|$ by p-stablity property.

  Apply Chernoff bounds.

Sketches and Space

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p-q)$ from $C(p)$ and $C(p)$

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p\text{-}q)$ from C(p) and C(p)

  CAUTION: Not an embedding into a normed space.

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p-q)$ from $C(p)$ and $C(p)$

  CAUTION: Not an embedding into a normed space.

- *Can we also construct sketch in small space:*

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p\text{-}q)$ from $C(p)$ and $C(p)$

  CAUTION: Not an embedding into a normed space.

- *Can we also construct sketch in small space:*

  Storing all x^i requires $\Omega(nk)$ space.

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p-q)$ from $C(p)$ and $C(p)$

  CAUTION: Not an embedding into a normed space.

- *Can we also construct sketch in small space:*

  Storing all x^i requires $\Omega(nk)$ space.

  Generate x^i with Nisan's pseudo-random generator.

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p-q)$ from $C(p)$ and $C(p)$

  CAUTION: Not an embedding into a normed space.

- *Can we also construct sketch in small space:*

  Storing all x^i requires $\Omega(nk)$ space.

  Generate x^i with Nisan's pseudo-random generator.

  Can store the seed in O(polylog n) space.

Sketches and Space

- *Sketch/Embedding into Small Dimension:*

  Let $x^1, x^2, \ldots, x^k$ be length n vector with $x_j^i \sim \mu$

  Let $C(y) = (x^1.y, \ldots, x^k.y)$

  Approximate $L_1(p\text{-}q)$ from $C(p)$ and $C(p)$

  CAUTION: Not an embedding into a normed space.

- *Can we also construct sketch in small space:*

  Storing all x^i requires $\Omega(nk)$ space.

  Generate x^i with Nisan's pseudo-random generator.

  Can store the seed in O(polylog n) space.

- *Thm:* Can $(1+\epsilon)$-approx $L_p(p\text{-}q)$ in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space.

1. Sketching L_p distances

p-stable distributions, pseudo-random generators

2. The Unsketchables

information divergences, communication complexity

3. Sketching Sketches

identifying correlations in data streams

Results

Results

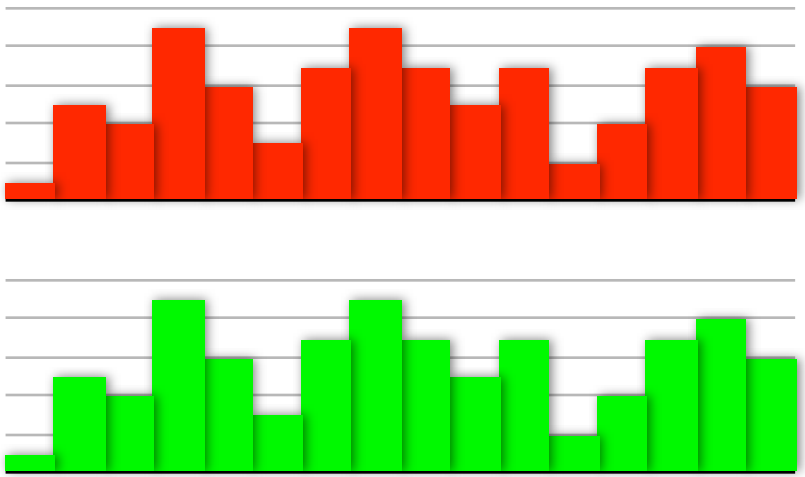
- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

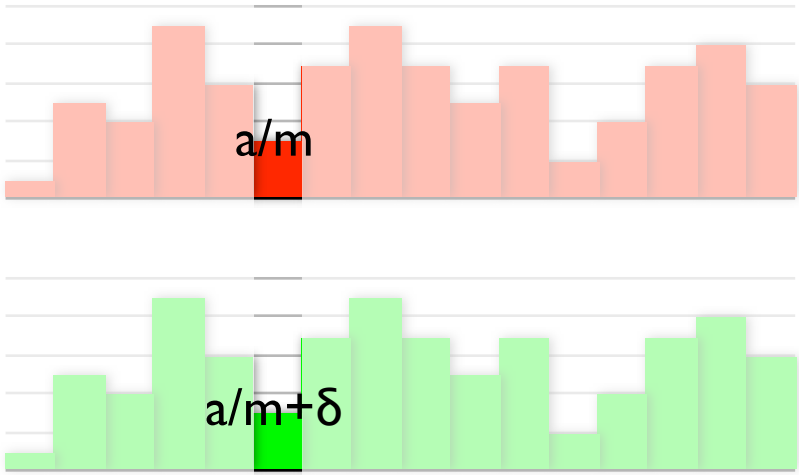
$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$



Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

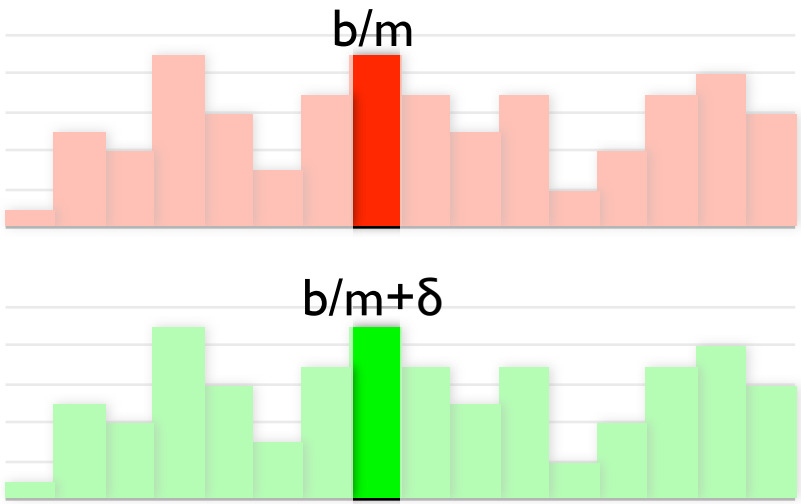
$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$



Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$



Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

- *Corollary:* Poly(n)-approx. of D_f requires $\Omega(n)$ space if f is twice differentiable and strictly convex.

Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

- *Corollary:* Poly(n)-approx. of D_f requires $\Omega(n)$ space if f is twice differentiable and strictly convex.

- *Corollary:* Poly(n)-approx. of B_F requires $\Omega(n)$ space if there exists ρ, z_0 with

$$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \ge (z_1/z_2)^\rho$$

or $$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \le (z_2/z_1)^\rho$$

Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

- *Corollary:* Poly(n)-approx. of D_f requires $\Omega(n)$ space if f is twice differentiable and strictly convex.

- *Corollary:* Poly(n)-approx. of B_F requires $\Omega(n)$ space if there exists ρ, z_0 with

$$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \ge (z_1/z_2)^\rho$$

or $$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \le (z_2/z_1)^\rho$$

- Only exceptions are L_1 and L_2!

Results

- *Thm (Shift Invariance):* t-approx. of $\sum \phi(p_i, q_i)$ needs $\Omega(n)$ space if for some a, b, c and $m=an/4+bn+cn/2$,

$$\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) > \frac{t^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

- *Corollary:* Poly(n)-approx. of D_f requires $\Omega(n)$ space if f is twice differentiable and strictly convex.

- *Corollary:* Poly(n)-approx. of B_F requires $\Omega(n)$ space if there exists ρ, z_0 with

$$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \ge (z_1/z_2)^\rho$$

or $$\forall\ 0 \le z_2 \le z_1 \le z_0,\ F''(z_1)/F''(z_2) \le (z_2/z_1)^\rho$$

- Only exceptions are L_1 and L_2!

! BREAKING NEWS: Many of these lower bounds also apply for randomly ordered streams [Chakrabarti, Cormode, McGregor 2007]

Alice

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

weight $n/4$

Bob

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

weight $n/4$

Alice

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

weight $n/4$

Question:

Are x and y disjoint, *i.e.*, $x.y=0$?

Thm (Razborov '92):

Needs $\Omega(n)$ communication.

Bob

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

weight $n/4$

$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

Alice

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

weight $n/4$

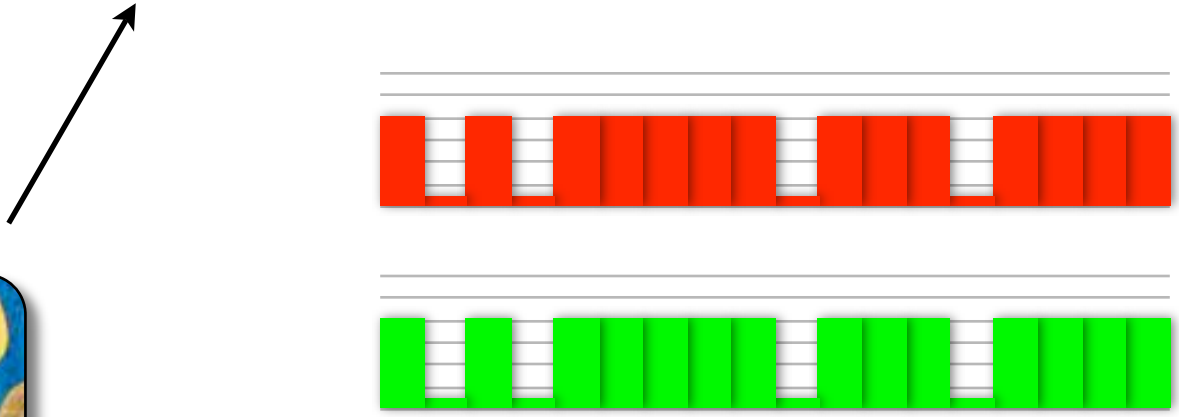
Bob

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

weight $n/4$

$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)



Alice

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

weight $n/4$

Bob

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

weight $n/4$

$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

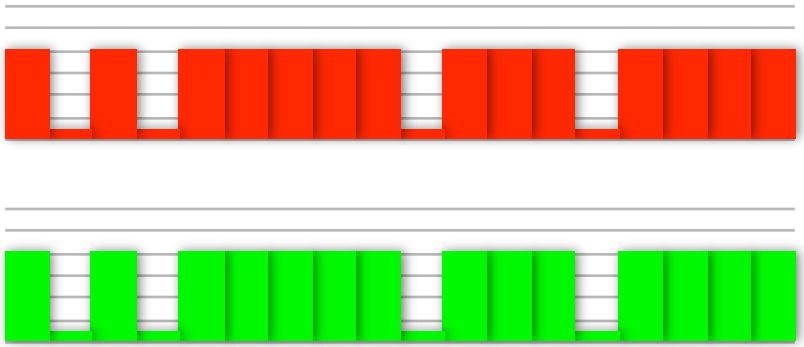
cy_i copies of i ($i \in [n]$)

c copies of $i+n$ ($i \in [n/4]$)

Alice

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

weight $n/4$



Bob

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

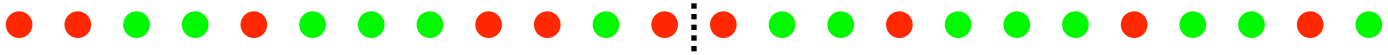
weight $n/4$

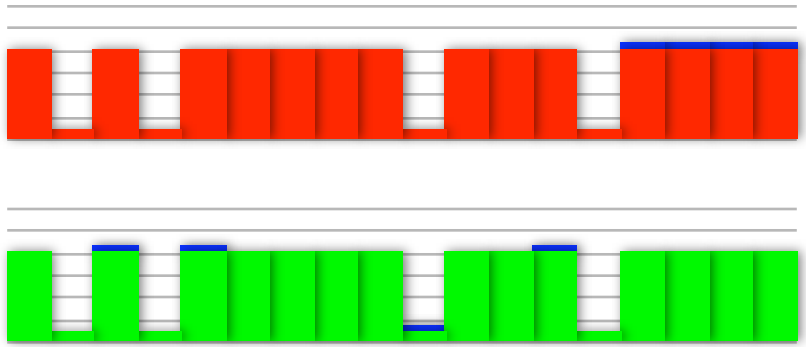
$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

cy_i copies of i ($i \in [n]$)

c copies of $i+n$ ($i \in [n/4]$)





Alice

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

weight $n/4$

Bob

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

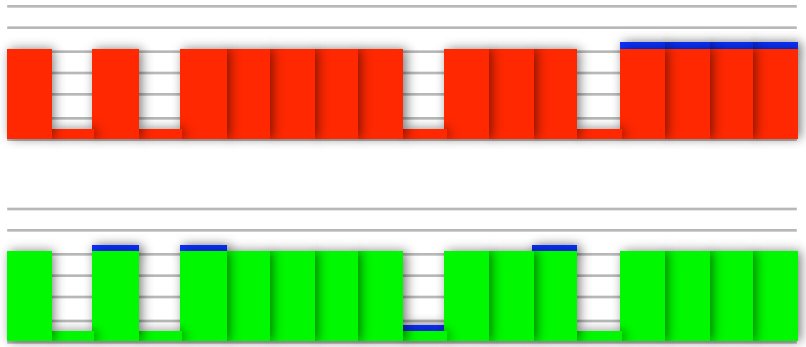
weight $n/4$

$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

cy_i copies of i ($i \in [n]$)

c copies of $i+n$ ($i \in [n/4]$)



Alice

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

weight $n/4$

Bob

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

weight $n/4$

If $x.y = 0$ then divergence is

$$\frac{n}{4}\left(\phi\left(\frac{b}{m},\frac{b+c}{m}\right)+\phi\left(\frac{b+c}{m},\frac{b}{m}\right)\right)$$

If $x.y = 1$ then divergence is at least

$$\phi\left(\frac{a}{m},\frac{a+c}{m}\right)$$

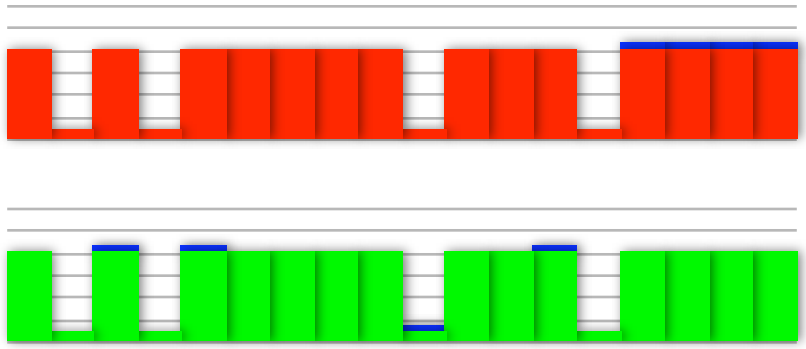
Factor t^2 difference by assumption

$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

cy_i copies of i ($i \in [n]$)

c copies of $i+n$ ($i \in [n/4]$)



Alice

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

weight $n/4$

Bob

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

weight $n/4$

Let $\mathscr{A}$ be a t-approx algorithm:

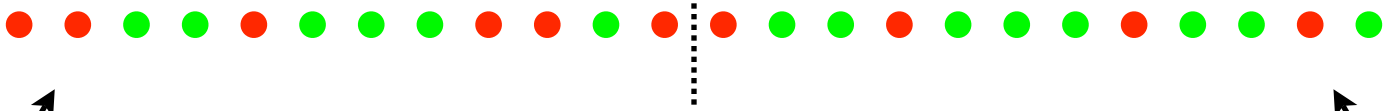
1. Alice runs $\mathscr{A}$ on first half
2. Transmits memory state
3. Bob instantiates $\mathscr{A}$
4. Continues $\mathscr{A}$ on second half

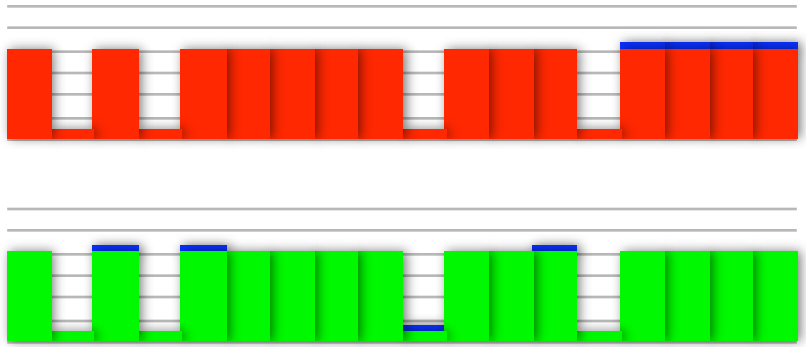
$ax_i+b(1-x_i)$ copies of i & i ($i \in [n]$)

b copies of $i+n$ & $i+n$ ($i \in [n/4]$)

cy_i copies of i ($i \in [n]$)

c copies of $i+n$ ($i \in [n/4]$)





Alice

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

weight $n/4$

Bob

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

weight $n/4$

Let $\mathscr{A}$ be a t-approx algorithm:

1. Alice runs $\mathscr{A}$ on first half
2. Transmits memory state
3. Bob instantiates $\mathscr{A}$
4. Continues $\mathscr{A}$ on second half

Thm: Any t-approx algorithm for the divergence requires $\Omega(n)$ memory.

Corollary to D_f

Corollary to D_f

- *Corollary:* Any *poly*(n) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if f'' exists and is strictly positive.

Corollary to D_f

- *Corollary:* Any *poly*(n) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if f'' exists and is strictly positive.
- *Proof:* Set a=c=1 and b=$t^2 n(f'(1)+1)/\ 8f(2)$

Corollary to D_f

- *Corollary:* Any *poly*(*n*) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if *f*" exists and is strictly positive.
- *Proof:* Set *a*=*c*=1 and $b=t^2n(f'(1)+1)/\ 8f(2)$

  Take Taylor expansion of *f* around 1:

$$\phi(b/m, (b+c)/m) \quad = \quad (b/m)\left[f(1) + f'(1)/b + f''(1+\gamma)/(2b^2)\right]$$

Corollary to D_f

- *Corollary:* Any *poly*(*n*) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if *f*" exists and is strictly positive.
- *Proof:* Set *a*=*c*=1 and $b=t^2n(f'(1)+1)/\ 8f(2)$

  Take Taylor expansion of *f* around 1:

$$\begin{aligned}\phi(b/m,(b+c)/m) &= (b/m)\left[f(1)+f'(1)/b+f''(1+\gamma)/(2b^2)\right]\\ &\leq (f''(1)+1)/(2mb)\end{aligned}$$

Corollary to D_f

- *Corollary:* Any *poly*(*n*) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if *f*" exists and is strictly positive.
- *Proof:* Set *a*=*c*=1 and *b*=$t^2 n(f'(1)+1)/ 8f(2)$

  Take Taylor expansion of *f* around 1:

$$\begin{aligned}\phi(b/m, (b+c)/m) &= (b/m)\left[f(1) + f'(1)/b + f''(1+\gamma)/(2b^2)\right] \\ &\leq (f''(1)+1)/(2mb) \\ &\leq 8\phi(a/m, (a+c)/m)/(t^2 n)\end{aligned}$$

Corollary to D_f

- *Corollary:* Any *poly*(*n*) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if *f*'' exists and is strictly positive.
- *Proof:* Set *a*=*c*=1 and $b=t^2n(f'(1)+1)/\ 8f(2)$

  Take Taylor expansion of *f* around 1:

$$\begin{aligned}\phi(b/m,(b+c)/m) &= (b/m)\left[f(1)+f'(1)/b+f''(1+\gamma)/(2b^2)\right]\\ &\leq (f''(1)+1)/(2mb)\\ &\leq 8\phi(a/m,(a+c)/m)/(t^2n)\end{aligned}$$

  Similarly, $\phi((b+c)/m, b/m) \leq 8\phi(a/m,(a+c)/m)/(t^2n)$

Corollary to D_f

- *Corollary:* Any *poly*(*n*) approx. of $D_f(p, q) = \sum p_i f(q_i/p_i)$ requires $\Omega(n)$ space if *f*" exists and is strictly positive.
- *Proof:* Set *a*=*c*=1 and $b=t^2n(f'(1)+1)/\ 8f(2)$

  Take Taylor expansion of *f* around 1:

$$\begin{aligned}\phi(b/m,(b+c)/m) &= (b/m)\left[f(1)+f'(1)/b+f''(1+\gamma)/(2b^2)\right]\\ &\leq (f''(1)+1)/(2mb)\\ &\leq 8\phi(a/m,(a+c)/m)/(t^2n)\end{aligned}$$

  Similarly, $\phi((b+c)/m, b/m) \leq 8\phi(a/m,(a+c)/m)/(t^2n)$

  Result follows by the Shift-Invariant Theorem.

Additive Error Algorithms

Additive Error Algorithms

- *f-Divergences:*

  Thm: If D_f bounded, $O_\epsilon(\ln \delta^{-1})$ space is sufficient for $\pm\epsilon$ approx. with prob. $1-\delta$.

  Thm: If D_f unbounded, need $\Omega(n)$ space.

Additive Error Algorithms

- *f-Divergences:*

  Thm: If D_f bounded, $O_\epsilon(\ln \delta^{-1})$ space is sufficient for $\pm\epsilon$ approx. with prob. $1-\delta$.

  Thm: If D_f unbounded, need $\Omega(n)$ space.

- *Bregman Divergences:*

  Thm: If $F(0)$, $F'(0)$, and $F''(\cdot)$ exist, $O_\epsilon(\ln \delta^{-1})$ space is sufficient for $\pm\epsilon$ approx. with prob. $1-\delta$.

  Thm: If $F(0)$ or $F'(0)$ infinite, need $\Omega(n)$ space.

1. Sketching L_p distances

p-stable distributions, pseudo-random generators

2. The Unsketchables

information divergences, communication complexity

3. Sketching Sketches

identifying correlations in data streams

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

  X_q with distribution $(q_1, ..., q_n)$

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

  X_q with distribution $(q_1, ..., q_n)$

  (X_p, X_q) with distribution $(r_{11}, r_{12}, ..., r_{nn})$

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

  X_q with distribution $(q_1, ..., q_n)$

  (X_p, X_q) with distribution $(r_{11}, r_{12}, ..., r_{nn})$

- *"How independent are X_p and X_q?"*

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

  X_q with distribution $(q_1, ..., q_n)$

  (X_p, X_q) with distribution $(r_{11}, r_{12}, ..., r_{nn})$

- *"How independent are X_p and X_q?"*

  Is the joint distribution "far" from the product distribution $(s_{11}, s_{12}, ..., s_{nn})$?

New Problem

- List of m pairs in $[n] \times [n]$:

  (3,5), (5,3), (2,7), (3,4), (7,1), (1,2), (3,9), (6,6), ...

- Stream defines random variables:

  X_p with distribution $(p_1, ..., p_n)$

  X_q with distribution $(q_1, ..., q_n)$

  (X_p, X_q) with distribution $(r_{11}, r_{12}, ..., r_{nn})$

- *"How independent are X_p and X_q?"*

  Is the joint distribution "far" from the product distribution $(s_{11}, s_{12}, ..., s_{nn})$?

  Consider $L_1(s\text{-}r)$ or $L_2(s\text{-}r)$ or mutual information:

$$I(X_p; X_q) = H(X_p) + H(X_q) - H(X_p, X_q) = \sum_{i,j} r_{ij} \lg \frac{p_i}{r_{ij}}$$

Results

- *Estimating $L_2(s\text{-}r)$:*

  Thm: $(1+\epsilon)$-factor approx. (w/p $1-\delta$) in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space.

- *Estimating $L_1(s\text{-}r)$:*

  Thm: $O(\ln n)$-factor approx. (w/p $1-\delta$) in $\tilde{O}(\ln \delta^{-1})$ space.

  Thm: $\pm\epsilon$ approx. (w/p $1-\delta$) in $\tilde{O}(\epsilon^{-4} \ln \delta^{-1})$ space (2-pass).

- *Estimating $I(X_p, X_q)$:*

  Thm: No 5/4-factor approx. (w/p 4/5) in $O(n)$ space.

  Thm: $\pm\epsilon$ approx. (w/p $1-\delta$) in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space.

L_2 Sketching Revisited

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.
- Compute x.(p-q)....

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.
- Compute x.(p-q)....

$$E[(x.(p-q))^2] \quad = \quad \Sigma_{i,j} E[x_i x_j](p_i - q_i)(p_j - q_j)$$

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.
- Compute x.(p-q)....

$$\mathrm{E}[(x.(p-q))^2] \quad = \quad \Sigma_{i,j}\mathrm{E}[x_i x_j](p_i - q_i)(p_j - q_j) = (L_2(p-q))^2$$

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.
- Compute x.(p-q)....

$$\begin{aligned} \mathrm{E}[(x.(p-q))^2] &= \Sigma_{i,j}\mathrm{E}[x_i x_j](p_i-q_i)(p_j-q_j) = (L_2(p-q))^2 \\ \\ \mathrm{Var}[(x.(p-q))^2] &\leq \mathrm{E}[(x.(p-q))^4] \\ &= \Sigma_{i,j,k,l}\mathrm{E}[x_i x_j x_k x_l](p_i-q_i)(p_j-q_j)(p_k-q_k)(p_l-q_l) \end{aligned}$$

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.

- Compute x.(p-q)....

$$\mathrm{E}[(x.(p-q))^2] = \Sigma_{i,j}\mathrm{E}[x_i x_j](p_i - q_i)(p_j - q_j) = (L_2(p-q))^2$$

$$\begin{aligned}\mathrm{Var}[(x.(p-q))^2] &\leq \mathrm{E}[(x.(p-q))^4] \\ &= \Sigma_{i,j,k,l}\mathrm{E}[x_i x_j x_k x_l](p_i - q_i)(p_j - q_j)(p_k - q_k)(p_l - q_l) \\ &= (L_2(p-q))^4\end{aligned}$$

L_2 Sketching Revisited

- Let $x \in \{-1,1\}^n$ where x_i are unbiased *4-wise indept*.

- Compute x.(p-q)....

$$\begin{aligned} \mathrm{E}[(x.(p-q))^2] &= \Sigma_{i,j}\mathrm{E}[x_i x_j](p_i-q_i)(p_j-q_j) = (L_2(p-q))^2 \\ \\ \mathrm{Var}[(x.(p-q))^2] &\leq \mathrm{E}[(x.(p-q))^4] \\ &= \Sigma_{i,j,k,l}\mathrm{E}[x_i x_j x_k x_l](p_i-q_i)(p_j-q_j)(p_k-q_k)(p_l-q_l) \\ &= (L_2(p-q))^4 \end{aligned}$$

- *Thm:* By Chebychev bounds, the average of $O(\epsilon^{-2} \ln \delta^{-1})$ repetitions yields $(1\pm\epsilon)$ $L_2(p\text{-}q)$ with prob. $1\text{-}\delta$.

[Alon, Matias, Szegedy 1996]

Testing L_2 Independence

Testing L_2 Independence

- *Idea:* Estimate L_2(*r*-*s*) using $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.

Testing L_2 Independence

- *Idea:* Estimate L_2(*r*-*s*) using $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.
- *Problem:* Can't compute sketch of product distribution!

Testing L_2 Independence

- *Idea:* Estimate L_2(r-s) using $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.
- *Problem:* Can't compute sketch of product distribution!
- *Solution:* Let $x, y \in \{-1, 1\}^n$ be 4-wise independent and set $z_{ij} = x_i y_j$.

$$z.s = \sum_{ij} z_{ij} s_{ij} = (x.p)(y.q)$$

Testing L_2 Independence

- *Idea:* Estimate L_2(r-s) using $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.
- *Problem:* Can't compute sketch of product distribution!
- *Solution:* Let $x, y \in \{-1, 1\}^n$ be 4-wise independent and set $z_{ij} = x_i y_j$.

$$z.s = \sum_{ij} z_{ij} s_{ij} = (x.p)(y.q)$$

- Entries are no longer 4-wise independent but it's okay. Let $a_{ij} = r_{ij} - s_{ij}$, and consider $T = \Sigma_{ij}\, z_{ij}\, a_{ij}$:

$$\begin{aligned} \mathrm{Var}[T^2] &\leq \mathrm{E}[T^4] \\ &= \Sigma_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4} \mathrm{E}[z_{i_1 j_1} z_{i_2 j_2} z_{i_3 j_3} z_{i_4 j_4}] a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4} \end{aligned}$$

Testing L_2 Independence

- *Idea:* Estimate L_2(*r*-*s*) using $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.
- *Problem:* Can't compute sketch of product distribution!
- *Solution:* Let $x, y \in \{-1, 1\}^n$ be 4-wise independent and set $z_{ij} = x_i y_j$.

$$z.s = \sum_{ij} z_{ij} s_{ij} = (x.p)(y.q)$$

- Entries are no longer 4-wise independent but it's okay. Let $a_{ij} = r_{ij} - s_{ij}$, and consider $T = \Sigma_{ij}\, z_{ij}\, a_{ij}$:

$$\begin{aligned} \mathrm{Var}[T^2] &\leq \mathrm{E}[T^4] \\ &= \Sigma_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4} \mathrm{E}[z_{i_1 j_1} z_{i_2 j_2} z_{i_3 j_3} z_{i_4 j_4}] a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4} \\ &\leq 3(L_2(p-q))^4 \end{aligned}$$

Testing L_2 Independence

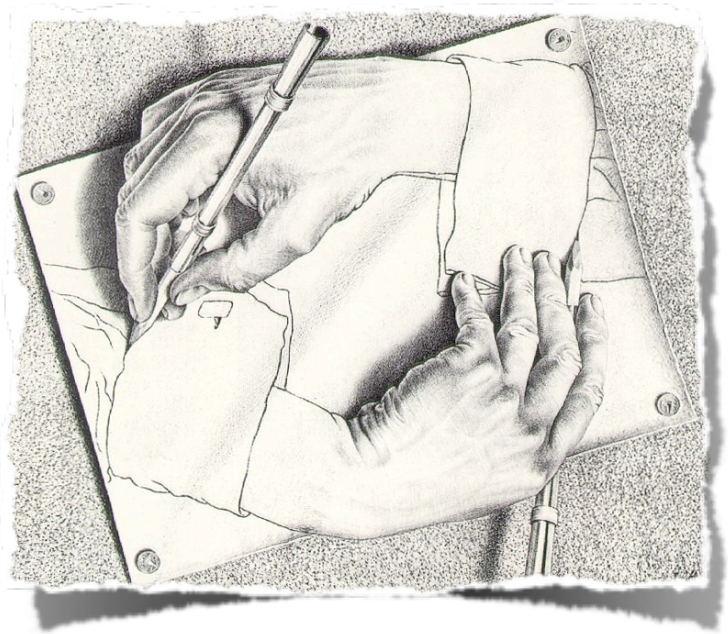
- *Idea:* Estimate L_2(*r*-*s*) using $z \in \{-1,1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent.
- *Problem:* Can't compute sketch of product distribution!
- *Solution:* Let $x, y \in \{-1,1\}^n$ be 4-wise independent and set $z_{ij} = x_i y_j$.

$$z.s = \sum_{ij} z_{ij} s_{ij} = (x.p)(y.q)$$

- Entries are no longer 4-wise independent but it's okay. Let $a_{ij} = r_{ij} - s_{ij}$, and consider $T = \Sigma_{ij}\, z_{ij}\, a_{ij}$:

$$\begin{aligned} \mathrm{Var}[T^2] &\leq \mathrm{E}[T^4] \\ &= \Sigma_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4} \mathrm{E}[z_{i_1 j_1} z_{i_2 j_2} z_{i_3 j_3} z_{i_4 j_4}] a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4} \\ &\leq 3(L_2(p-q))^4 \end{aligned}$$

- Repeat $O(\epsilon^{-2} \ln \delta^{-1})$ times to deduce $(1 \pm \epsilon)$ L_2(*r*-*s*)



Summary

Small space sketches of L_1 and L_2 using p-stable distributions.

No small space sketches exists for other information divergences.

Can use sketching ideas to estimate independence.

Main Material:

Stable distributions, pseudo-random generators, embeddings, and data stream computation
Piotr Indyk (FOCS 2000)

Sketching information divergences
Sudipto Guha, Piotr Indyk, Andrew McGregor (COLT 2007)

Declaring independence via the sketching of sketches
Piotr Indyk, Andrew McGregor (SODA 2008)