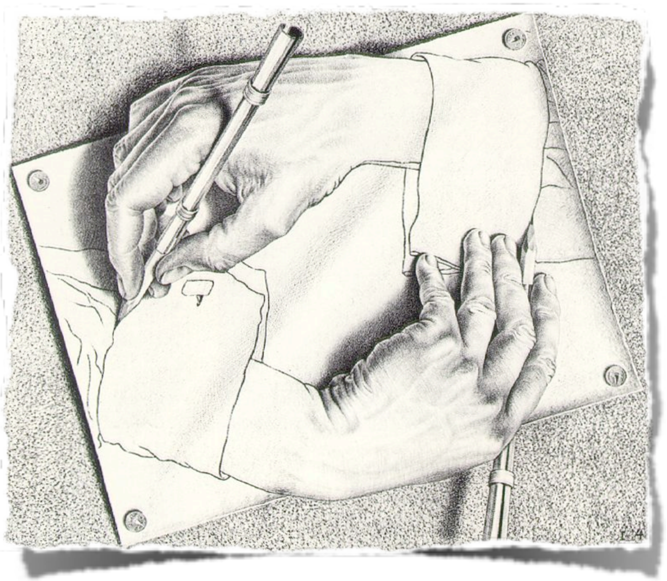
Declaring Independence via the Sketching of Sketches

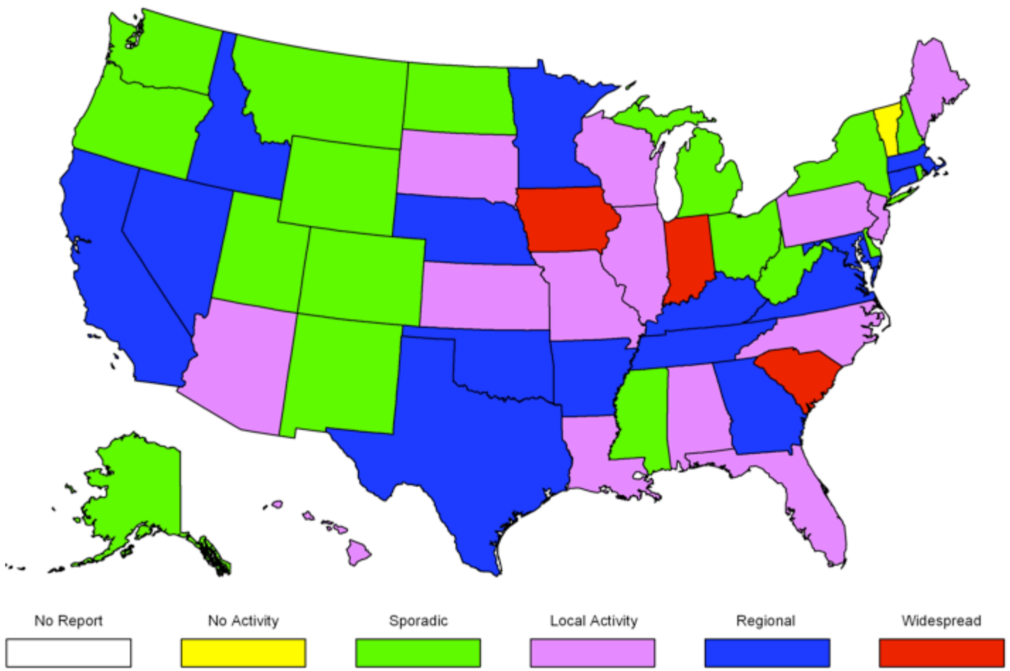


Piotr Indyk
Massachusetts Institute of Technology

Andrew McGregor
University of California, San Diego
Until August '08 -- Hire Me!

The Problem

The Problem

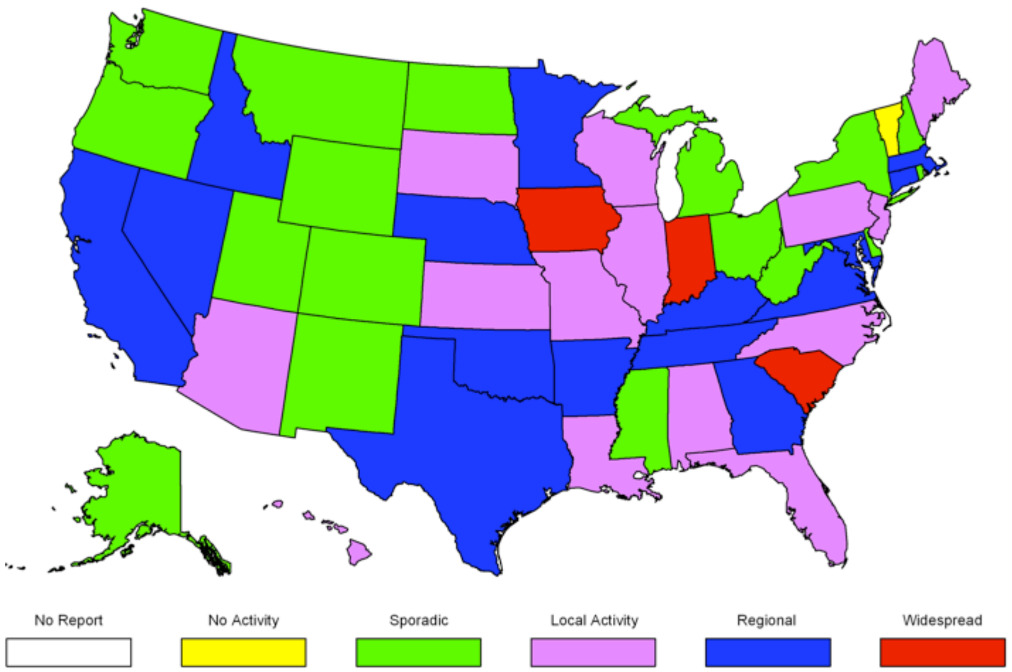


Center for Disease Control (CDC) has massive amounts of data on disease occurrences and their locations.

"How correlated is your zip code to the diseases you'll catch this year?"

Image from http://www.cdc.gov/flu/weekly/weeklyarchives2006-2007/images/usmap02.jpg

The Problem



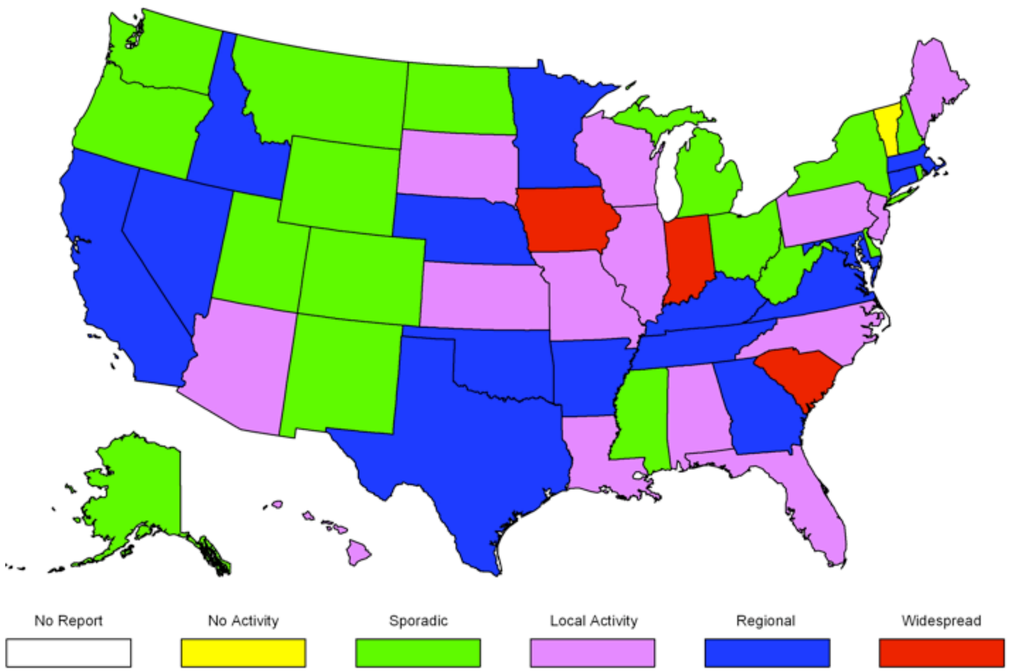
Center for Disease Control (CDC) has massive amounts of data on disease occurrences and their locations.

"How correlated is your zip code to the diseases you'll catch this year?"

- *Sample (sub-linear time):*
  How many are required to distinguish independence from "ϵ-far" from independence? [Batu et al. '01], [Alon et al. '07], [Valiant '08]

Image from http://www.cdc.gov/flu/weekly/weeklyarchives2006-2007/images/usmap02.jpg

The Problem



Center for Disease Control (CDC) has massive amounts of data on disease occurrences and their locations.

"How correlated is your zip code to the diseases you'll catch this year?"

- *Sample (sub-linear time):*
  How many are required to distinguish independence from "ϵ-far" from independence? [Batu et al. '01], [Alon et al. '07], [Valiant '08]
- *Stream (sub-linear space):*
  Access pairs sequentially or "online" and limited memory.

Image from http://www.cdc.gov/flu/weekly/weeklyarchives2006-2007/images/usmap02.jpg

Formulation

Formulation

- Stream of *m* pairs in [*n*] x [*n*]:

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

Formulation

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

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

- Define "empirical" distributions:

  Marginals: $(p_1, \ldots, p_n)$, $(q_1, \ldots, q_n)$

  Joint: $(r_{11}, r_{12}, \ldots, r_{nn})$

  Product: $(s_{11}, s_{12}, \ldots, s_{nn})$ where s_{ij} *equals* $p_i\, q_j$

Formulation

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

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

- Define "empirical" distributions:

  Marginals: $(p_1, \ldots, p_n)$, $(q_1, \ldots, q_n)$

  Joint: $(r_{11}, r_{12}, \ldots, r_{nn})$

  Product: $(s_{11}, s_{12}, \ldots, s_{nn})$ where s_{ij} *equals* $p_i\, q_j$

- *Question:* How correlated are first and second terms?

Formulation

- Stream of m pairs in $[n]$ x $[n]$:

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

- Define "empirical" distributions:

  Marginals: $(p_1, ..., p_n)$, $(q_1, ..., q_n)$

  Joint: $(r_{11}, r_{12}, ..., r_{nn})$

  Product: $(s_{11}, s_{12}, ..., s_{nn})$ where s_{ij} *equals* $p_i\, q_j$

- *Question:* How correlated are first and second terms?

  E.g.,

$$L_1(s-r) = \sum_{i,j} |s_{ij} - r_{ij}|$$

$$L_2(s-r) = \sqrt{\sum_{i,j} (s_{ij} - r_{ij})^2}$$

$$I(s,r) = H(p) - H(p|q)$$

Formulation

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

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

- Define "empirical" distributions:

  Marginals: $(p_1, \ldots, p_n), (q_1, \ldots, q_n)$

  Joint: $(r_{11}, r_{12}, \ldots, r_{nn})$

  Product: $(s_{11}, s_{12}, \ldots, s_{nn})$ where s_{ij} *equals* $p_i\, q_j$

- *Question:* How correlated are first and second terms?

  E.g.,
  $$L_1(s-r) = \sum_{i,j} |s_{ij} - r_{ij}|$$
  $$L_2(s-r) = \sqrt{\sum_{i,j} (s_{ij} - r_{ij})^2}$$
  $$I(s,r) = H(p) - H(p|q)$$

- *Previous work:* Can estimate L_1 *and* L_2 between marginals.

  [Alon, Matias, Szegedy '96], [Feigenbaum et al. '99], [Indyk '00], [Guha, Indyk, McGregor '07], [Ganguly, Cormode '07]

Our Results

Our Results

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

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

  "Neat" result extending AMS sketches

Our Results

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

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

  "Neat" result extending AMS sketches

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

  $O(\ln n)$-factor approx. in $\tilde{O}(\ln \delta^{-1})$ space.

  Sketches of sketches and sketches/embeddings

Our Results

- ***Estimating $L_2(s-r)$:***

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

  "Neat" result extending AMS sketches

- ***Estimating $L_1(s-r)$:***

  $O(\ln n)$-factor approx. in $\tilde{O}(\ln \delta^{-1})$ space.

  Sketches of sketches and sketches/embeddings

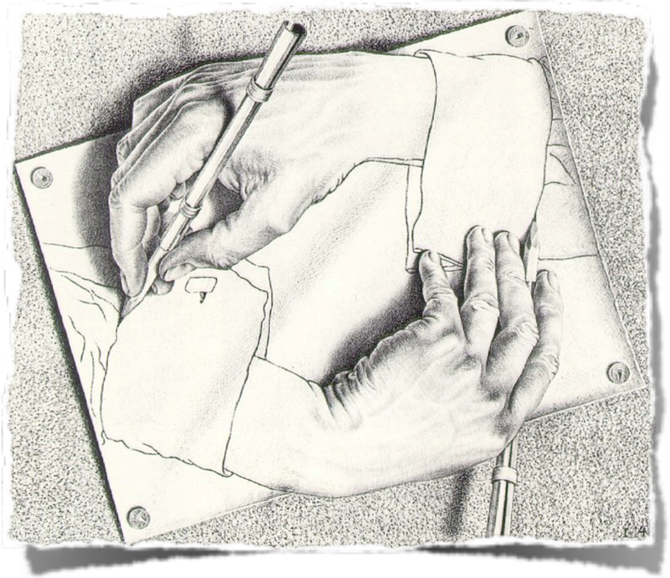
- ***Other Results:***

  $L_1(s-r)$: Additive approximations

  Mutual Information: Additive but not $(1+\epsilon)$-factor approx.

  Distributed Model: Pairs are observed by different parties.

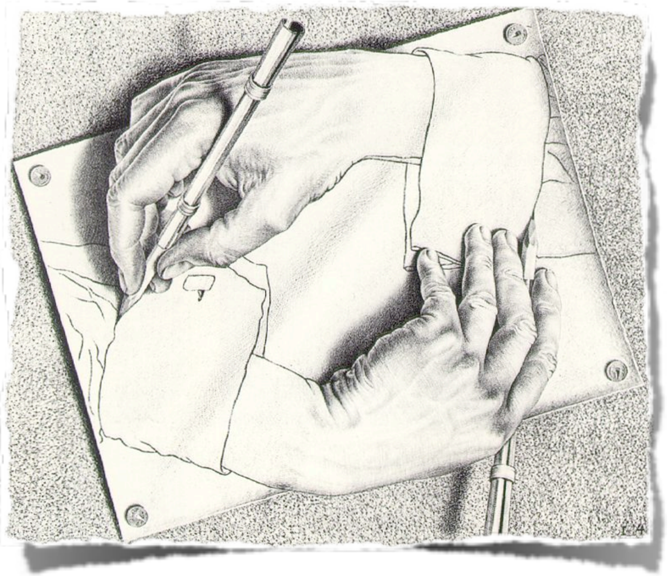
a) Neat Result for L_2
b) Sketching Sketches
c) Other Results



a) Neat Result for L_2

b) Sketching Sketches

c) Other Results



First Attempt

First Attempt

- *Random Projection:* Let $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]

First Attempt

- *Random Projection:* Let $z \in \{-1,1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]
- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

First Attempt

- *Random Projection:* Let $z \in \{-1,1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]
- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

- Correct in expectation and has small variance:

$$\mathsf{E}[T] = \Sigma_{i_1,j_1,i_2,j_2} \mathsf{E}[z_{i_1j_1} z_{i_2j_2}] a_{i_1j_1} a_{i_2j_2} = (L_2(r-s))^2$$

$$(a_{ij} = r_{ij} - s_{ij})$$

$$\begin{aligned} \mathsf{Var}[T] &\leq \mathsf{E}[T^2] \\ &= \Sigma_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4} \mathsf{E}[z_{i_1j_1} z_{i_2j_2} z_{i_3j_3} z_{i_4j_4}] a_{i_1j_1} a_{i_2j_2} a_{i_3j_3} a_{i_4j_4} \\ &\leq \mathsf{E}[T]^2 \end{aligned}$$

First Attempt

- *Random Projection:* Let $z \in \{-1,1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]
- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

- Correct in expectation and has small variance:

$$\mathsf{E}[T] \quad = \quad \Sigma_{i_1,j_1,i_2,j_2} \mathsf{E}[z_{i_1 j_1} z_{i_2 j_2}] a_{i_1 j_1} a_{i_2 j_2} = (L_2(r-s))^2$$

$$(a_{ij} = r_{ij} - s_{ij})$$

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

First Attempt

- *Random Projection:* Let $z \in \{-1, 1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]

- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

- Correct in expectation and has small variance:

$$\mathrm{E}[T] = \Sigma_{i_1,j_1,i_2,j_2} \mathrm{E}[z_{i_1 j_1} z_{i_2 j_2}] a_{i_1 j_1} a_{i_2 j_2} = (L_2(r - s))^2$$

$$(a_{ij} = r_{ij} - s_{ij})$$

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

First Attempt

- *Random Projection:* Let $z \in \{-1,1\}^{n \times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]
- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

- Correct in expectation and has small variance:

$$\mathsf{E}[T] = \Sigma_{i_1,j_1,i_2,j_2} \mathsf{E}[z_{i_1 j_1} z_{i_2 j_2}] a_{i_1 j_1} a_{i_2 j_2} = (L_2(r-s))^2$$

$$(a_{ij} = r_{ij} - s_{ij})$$

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

First Attempt

- *Random Projection:* Let $z \in \{-1,1\}^{n\times n}$ where z_{ij} are unbiased 4-wise independent. [Alon, Matias, Szegedy '96]
- *Estimator:* Suppose we can compute estimator:

$$T = (z.r - z.s)^2$$

- Correct in expectation and has small variance:

$$\mathsf{E}[T] = \Sigma_{i_1,j_1,i_2,j_2}\mathsf{E}[z_{i_1j_1}z_{i_2j_2}]a_{i_1j_1}a_{i_2j_2} = (L_2(r-s))^2$$

$$(a_{ij} = r_{ij} - s_{ij})$$

$$\begin{aligned}\mathsf{Var}[T] &\le \mathsf{E}[T^2] \\ &= \Sigma_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4}\mathsf{E}[z_{i_1j_1}z_{i_2j_2}z_{i_3j_3}z_{i_4j_4}]a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4} \\ &\le \mathsf{E}[T]^2\end{aligned}$$

- Repeating $O(\epsilon^{-2} \ln \delta^{-1})$ times and take the *mean*.

Computing Estimator

Computing Estimator

- Need to compute: $z.r$ and $z.s$

Computing Estimator

- Need to compute: $z.r$ and $z.s$
- <u>*Good News:*</u> First term is easy

```
1) Let A = 0
2) For each stream element:
   2.1) If stream element = (i,j) then A ← A + z_ij/m
```

Computing Estimator

- Need to compute: $z.r$ and $z.s$
- *Good News:* First term is easy

```
1) Let A = 0
2) For each stream element:
   2.1) If stream element = (i,j) then A ← A + z_ij/m
```

- *Bad News:* Can't compute second term!

Computing Estimator

- Need to compute: $z.r$ and $z.s$
- Good News: First term is easy

```
1) Let A = 0
2) For each stream element:
    2.1) If stream element = (i,j) then A ← A + zij/m
```

- Bad News: Can't compute second term!
- Good News: Use bilinear sketch: If $z_{ij} = x_i y_j$ for $x, y \in \{-1, 1\}^n$

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

i.e., product of sketches is sketch of product.

Computing Estimator

- Need to compute: $z.r$ and $z.s$
- *Good News:* First term is easy

```
1) Let A = 0
2) For each stream element:
   2.1) If stream element = (i,j) then A ←A + zij/m
```

- *Bad News:* Can't compute second term!
- *Good News:* Use bilinear sketch: If $z_{ij} = x_i y_j$ for $x, y \in \{-1, 1\}^n$

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

  i.e., product of sketches is sketch of product.

- *Bad News:* z is no longer 4-wise independent even if *x* and *y* are fully random, e.g.,

$$z_{11} z_{12} z_{21} z_{22} = (x_1)^2 (x_2)^2 (y_1)^2 (y_2)^2 = 1$$

Still Get Low Variance

Still Get Low Variance

- *Lemma:* Variance has at most tripled.

Still Get Low Variance

- *Lemma:* Variance has at most tripled.
- *Proof:*

$$z = \begin{pmatrix} x_1 y_1 & x_2 y_1 & \cdots & \cdots & x_n y_1 \\ x_1 y_2 & x_2 y_2 & \cdots & \cdots & x_n y_2 \\ \vdots & \vdots & & & \vdots \\ x_1 y_n & x_2 y_n & \cdots & \cdots & x_n y_n \end{pmatrix}$$

Still Get Low Variance

- *Lemma:* Variance has at most tripled.
- *Proof*:

$$z = \begin{pmatrix} x_1y_1 & x_2y_1 & \cdots & \cdots & x_ny_1 \\ x_1y_2 & x_2y_2 & \cdots & \cdots & x_ny_2 \\ \vdots & \vdots & & & \vdots \\ x_1y_n & x_2y_n & \cdots & \cdots & x_ny_n \end{pmatrix}$$

- Product of four entries is biased iff entries lie in rectangle

Still Get Low Variance

- *Lemma:* Variance has at most tripled.
- *Proof:*

$$z = \begin{pmatrix} x_1 y_1 & x_2 y_1 & \cdots & \cdots & x_n y_1 \\ x_1 y_2 & x_2 y_2 & \cdots & \cdots & x_n y_2 \\ \vdots & \vdots & & & \vdots \\ x_1 y_n & x_2 y_n & \cdots & \cdots & x_n y_n \end{pmatrix}$$

- Product of four entries is biased iff entries lie in rectangle
- Hence, $\mathrm{Var}[T] \leq \sum_{\substack{(i_1,j_1),(i_2,j_2),\\(i_3,j_3),(i_4,j_4)\\ \text{in rectangle}}} a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4}$

Still Get Low Variance

- *Lemma:* Variance has at most tripled.
- *Proof*:

$$z = \begin{pmatrix} x_1 y_1 & x_2 y_1 & \cdots & \cdots & x_n y_1 \\ x_1 y_2 & x_2 y_2 & \cdots & \cdots & x_n y_2 \\ \vdots & \vdots & & & \vdots \\ x_1 y_n & x_2 y_n & \cdots & \cdots & x_n y_n \end{pmatrix}$$

- Product of four entries is biased iff entries lie in rectangle
- Hence, $\mathrm{Var}[T] \leq \sum_{\substack{(i_1,j_1),(i_2,j_2),\\(i_3,j_3),(i_4,j_4)\\ \text{in rectangle}}} a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4}$

  since a rectangle is uniquely specified by a diagonal and

$$2 a_{i_1 j_1} a_{i_2 j_2} a_{i_3 j_3} a_{i_4 j_4} \leq (a_{i_1 j_1} a_{i_2 j_2})^2 + (a_{i_3 j_3} a_{i_4 j_4})^2$$

Still Get Low Variance

- *Lemma:* Variance has at most tripled.
- *Proof*:

$$z = \begin{pmatrix} x_1y_1 & x_2y_1 & \cdots & \cdots & x_ny_1 \\ x_1y_2 & x_2y_2 & \cdots & \cdots & x_ny_2 \\ \vdots & \vdots & & & \vdots \\ x_1y_n & x_2y_n & \cdots & \cdots & x_ny_n \end{pmatrix}$$

- Product of four entries is biased iff entries lie in rectangle
- Hence, $\mathrm{Var}[T] \leq \sum_{\substack{(i_1,j_1),(i_2,j_2),\\(i_3,j_3),(i_4,j_4)\\ \text{in rectangle}}} a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4} \leq 3\mathrm{E}[T]^2$

  since a rectangle is uniquely specified by a diagonal and

$$2a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4} \leq (a_{i_1j_1}a_{i_2j_2})^2 + (a_{i_3j_3}a_{i_4j_4})^2$$

Still Get Low Variance

- ***Lemma:*** Variance has at most tripled.
- ***Proof:***

$$z = \begin{pmatrix} x_1y_1 & x_2y_1 & \cdots & \cdots & x_ny_1 \\ x_1y_2 & x_2y_2 & \cdots & \cdots & x_ny_2 \\ \vdots & \vdots & & & \vdots \\ x_1y_n & x_2y_n & \cdots & \cdots & x_ny_n \end{pmatrix}$$

- Product of four entries is biased iff entries lie in rectangle
- Hence, $\text{Var}[T] \leq \sum_{\substack{(i_1,j_1),(i_2,j_2),\\(i_3,j_3),(i_4,j_4)\\ \text{in rectangle}}} a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4} \leq 3\text{E}[T]^2$

  since a rectangle is uniquely specified by a diagonal and

$$2a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4} \leq (a_{i_1j_1}a_{i_2j_2})^2 + (a_{i_3j_3}a_{i_4j_4})^2$$

- Less independence useful for range-sums. [Rusu, Dobra '06]

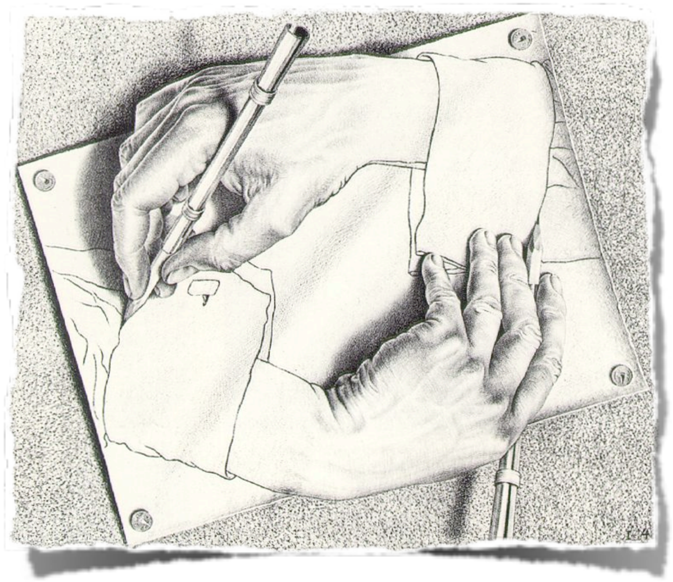
Summary of L_2 Result

- *Thm:* $(1+\epsilon)$-factor approx. (w/p $1-\delta$) in $\tilde{O}(\epsilon^{-2} \ln \delta^{-1})$ space.
- *Proof Ideas:*
  1) *First attempt:* Use AMS technique.
  2) *Road block:* Can't sketch product distribution.
  3) *Bilinear sketch:* Product of sketches was sketch of product!
  4) *PANIC:* No longer 4-wise independence.
  5) *Relax:* We didn't need full 4-wise independence.

a) Neat Result for L_2

***b)* Sketching Sketches**

c) Other Results



L_1 Result

L_1 Result

- ***Thm:*** O(ln *n*)-factor approx. of $L_1(s-r)$ in $\tilde{O}(\ln \delta^{-1})$ space.

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of $L_1(s-r)$ in $\tilde{O}(\ln \delta^{-1})$ space.
- Why not $(1+\epsilon)$-factor using Indyk's p-stable technique?

[Indyk, '00]

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of $L_1(s-r)$ in $\tilde{O}(\ln \delta^{-1})$ space.
- Why not $(1+\epsilon)$-factor using Indyk's p-stable technique?

[Indyk, '00]

- *Review of L_1 sketching:*

  Let entries of *z* be Cauchy(0,1)

  Compute estimator |z.a|

  Repeat $k=O(\epsilon^{-2} \ln \delta^{-1})$ times with different z.

  Take the *median* and appeal to concentration lemmas.

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of L_1(s-r) in $\tilde{O}(\ln \delta^{-1})$ space.
- Why not $(1+\epsilon)$-factor using Indyk's p-stable technique?

  [Indyk, '00]

- *Review of L_1 sketching:*

  Let entries of *z* be Cauchy(0,1)

  Compute estimator |z.a|

  Repeat $k=O(\epsilon^{-2} \ln \delta^{-1})$ times with different z.

  Take the *median* and appeal to concentration lemmas.

- *N.B.* If *median* were *mean* we'd have a dimensionality reduction result that doesn't exist. [Brinkman, Charikar '03]

Sketching Sketches

Sketching Sketches

- To sketch product distribution need $z = yM_x$

$$z = \underbrace{(\quad y \quad)}_{n} \underbrace{\begin{pmatrix} (\quad x \quad) & 0 & \ldots & 0 \\ 0 & (\quad x \quad) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & (\quad x \quad) \end{pmatrix}}_{n^2}$$

Sketching Sketches

- To sketch product distribution need $z = yM_x$

$$z = \underbrace{\begin{pmatrix} \quad y \quad \end{pmatrix}}_{n} \underbrace{\begin{pmatrix} \\ \qquad\qquad\qquad M_x \qquad\qquad\qquad \\ \\ \end{pmatrix}}_{n^2}$$

Sketching Sketches

- To sketch product distribution need $z = yM_x$

$$z = \underbrace{\begin{pmatrix} \quad y \quad \end{pmatrix}}_{n} \underbrace{\begin{pmatrix} \\ \qquad\qquad M_x \qquad\qquad \\ \\ \end{pmatrix}}_{n^2}$$

- *Sketch:* **Inner Sketch** **Outer Sketch**

$$\begin{array}{ccc} \mathbb{R}^{n^2} & \longmapsto & \mathbb{R}^n \\ a & \longrightarrow & M_x a \end{array} \qquad \begin{array}{ccc} \mathbb{R}^n & \longmapsto & \mathbb{R} \\ M_x a & \longrightarrow & yM_x a \end{array}$$

Sketching Sketches

- To sketch product distribution need $z = yM_x$

$$z = \underbrace{\left(\quad y \quad \right)}_{n} \underbrace{\left(\qquad\qquad M_x \qquad\qquad \right)}_{n^2}$$

- *Sketch:* *Inner Sketch* *Outer Sketch*

$$\begin{array}{ccc} \mathbb{R}^{n^2} & \longmapsto & \mathbb{R}^n \\ a & \longrightarrow & M_x a \end{array} \qquad \begin{array}{ccc} \mathbb{R}^{n} & \longmapsto & \mathbb{R} \\ M_x a & \longrightarrow & yM_x a \end{array}$$

- *The Problem:*

  Need to take median of multiple inner sketches before taking outer sketch.

Sketching Sketches

- To sketch product distribution need $z = yM_x$

$$z = \underbrace{\left(\quad y \quad \right)}_{n} \underbrace{\left(\qquad\qquad M_x \qquad\qquad \right)}_{n^2}$$

- *Sketch:* *Inner Sketch* *Outer Sketch*

$$\begin{array}{ccc} \mathbb{R}^{n^2} & \longmapsto & \mathbb{R}^n \\ a & \longrightarrow & M_x a \end{array} \qquad \begin{array}{ccc} \mathbb{R}^n & \longmapsto & \mathbb{R} \\ M_x a & \longrightarrow & yM_x a \end{array}$$

- *The Problem:*

  Need to take median of multiple inner sketches before taking outer sketch.

  The size of the inner sketch is large.

L_I Result

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of L_1(s-r) in Õ(ln δ^{-1}) space.

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of $L_1(s-r)$ in $\tilde{O}(\ln \delta^{-1})$ space.
- *Proof:*

  Outer sketch: Entries y are Cauchy(0,1)

  Inner sketch: Entries x are "truncated" Cauchy(0,1)

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of L_1(s-r) in Õ(ln δ^{-1}) space.
- *Proof:*

  Outer sketch: Entries y are Cauchy(0,1)

  Inner sketch: Entries x are "truncated" Cauchy(0,1)

$$\Pr\left[\Omega(1) \leq \frac{|M(x).a|}{|a|} \leq O(\log n)\right] \geq 9/10$$

L_1 Result

- *Thm:* O(ln *n*)-factor approx. of L_1(s-r) in Õ(ln δ^{-1}) space.
- *Proof:*

  Outer sketch: Entries y are Cauchy(0,1)

  Inner sketch: Entries x are "truncated" Cauchy(0,1)

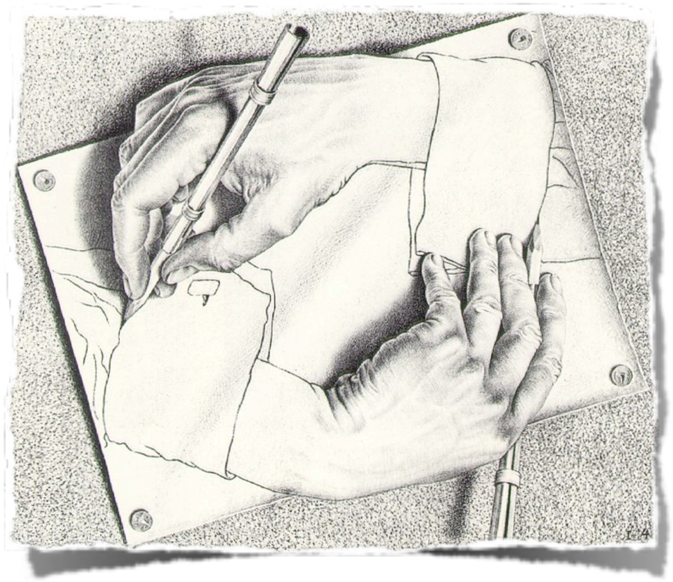
$$\Pr\left[\Omega(1) \leq \frac{|M(x).a|}{|a|} \leq O(\log n)\right] \geq 9/10$$

Repeat Õ(ln δ^{-1}) times and take median.

a) Neat Result for L_2

b) Sketching Sketches

c) Other Results



Other Results

Other Results

- *Mutual Information:*

  Can't $(1+\epsilon)$-factor approximate in $o(n)$ space

  Can $\pm\epsilon$ using algorithms for approx. entropy.

[Chakrabarti, Cormode, McGregor '07]

Other Results

- ***Mutual Information:***

  Can't $(1+\epsilon)$-factor approximate in $o(n)$ space

  Can $\pm\epsilon$ using algorithms for approx. entropy.

  [Chakrabarti, Cormode, McGregor '07]

- ***Distributed Model:***

  Player 1 sees $(3,\cdot), (5,\cdot), (2,\cdot), (3,\cdot), (7,\cdot), (1,\cdot), (3,\cdot), (6,\cdot), \ldots$

  Player 2 sees $(\cdot,5), (\cdot,3), (\cdot,7), (\cdot,4), (\cdot,1), (\cdot,2), (\cdot,9), (\cdot,6), \ldots$

  Very hard in general, e.g., can't check if $L_1(s-r)=0$

Other Results

- *Mutual Information:*

  Can't $(1+\epsilon)$-factor approximate in $o(n)$ space

  Can $\pm\epsilon$ using algorithms for approx. entropy.

  [Chakrabarti, Cormode, McGregor '07]

- *Distributed Model:*

  Player 1 sees (3,·), (5,·), (2,·), (3,·), (7,·), (1,·), (3,·), (6,·), ...

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

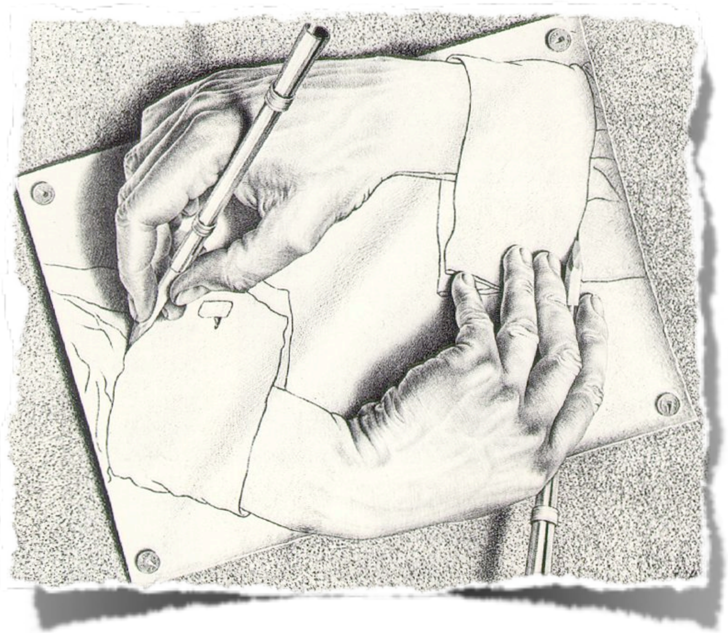
  Very hard in general, e.g., can't check if $L_1(s-r)=0$

- *Additive Approximation for $L_1(s-r)$:*

$$L_1(p-q) = \sum_i p_i L_1(q - q^i)$$

  where q^i is q conditioned on first term equals i.

  [Guha, McGregor, Venkatasubramanian '06]



Main Results

Can estimate $L_2(r-s)$ well using neat extension of AMS sketch.

Can estimate $L_1(r-s)$ up to $O(\log n)$ factor using p-stable distributions.

Can estimate mutual information additively using entropy algorithms.

Questions?