Zeroing in on the L_0 Metric

Graham Cormode

graham@dimacs.rutgers.edu

Joint work with M. Datar, P. Indyk.
S. Muthukrishnan

Oversensitivity

In some situations, L_1 and L_2 are too sensitive to the values in the vectors:

X:	6	5	4	3	2	1
Y:	100	5	4	3	2	1
Z:	12	12	12	12	12	12

Y looks more similar to X, but $\|X-Y\|_1 = 94$,
$\|X-Z\|_1 = 51$

Want a metric which is less sensitive to size of differences

Introducing L_0

Define L_0 norm (abusing terminology a little):

$$\|X\|_0 = \Sigma\, X_i^0$$

$$X_i = 0 \Rightarrow X_i^p = 0$$
$$X_i \neq 0 \Rightarrow X_i^p = 1$$

L_0 norm counts number of non-zero entries in the vector

Closely connected to F_0, zeroth Frequency Moment of a stream

L_0 Metric

L_0 Metric is

$$\|X\text{-}Y\|_0 = |\{i \mid X_i \neq Y_i\}|$$

Just count the number of locations where two vectors differ.

For strings this corresponds to the Hamming metric.

Uses of L_0

Many applications of L_0 norms and metric

- Machine Learning – used as an evaluation function
- Data Cleaning – find tables that are almost identical
- String Similarity (substitution errors)
- Clustering – less sensitive to outlier values
- Count number of distinct items in a stream

Embedding into L_1

L_0 metric on integer valued vectors trivially embeds into L_1 (equivalently, L_2^2 or L_p^p):

Let U(j) = bitstring that is all zeros, apart from 1 in j'th location

Set $U(X) = U(X_1)U(X_2)\ldots U(X_n)$

Then $\|X - Y\|_0 = \frac{1}{2} \|U(X) - U(Y)\|_1$

So, what's the problem?

Drawbacks

- The embedding only works in the sketch model – if the whole vector is available
- If a large vector is presented incrementally (in a streaming fashion with small space), can't make the embedding incrementally
- Example: tracking information about tables in databases, want to find near duplicates
- Tables are frequently updated with inserts and deletes, want to maintain information without rescanning

Incremental Model

- Can model most scenarios as a vector
- Example Vector a represents counts of each attribute value in a table
- Vectors defined incrementally by *updates* (i,j)
- (i,j) means add j to entry i in the (initially zero) vector a

Goal

Given vectors a, b, c… presented in this incremental format, want to be able to maintain a small space summary (sketch) of each vector.

Use this sketch to approximate (with guarantees):

- L_0 norm of any vector: $\|a\|_0$, $\|b\|_0$, …
- L_0 difference of any pair: $\|a - b\|_0$, $\|a - c\|_0$ …
- Union size of subset: $\|a + b + c\|_0$
- Arbitrary combinations: $\|a + b - c + \ldots\|_0$

Connection to F_0

- Suppose all updates are of the form (i,1)
- Then $\|a\|_0$ is number of distinct i's seen in stream = F_o, zeroth moment of the stream
- $F_0 = \|a\|_0$ if j positive for all updates (i,j)
- But not if j's are allowed to be negative – eg if we are computing $\|a - b\|_0$
- Many algorithms for F_0 (Flajolet, Martin 83, Bar-Yossef, Jayram, Kumar, Sivakumar 02, Gibbons, Tirthapura 01 and more) but can't handle negative updates.

Solution Method

- Sketch will be a linear projection of the vector with vectors with entries from appropriately chosen distributions
- Will show how to maintain this projection in small space
- Will show which distribution we want to draw values from, and how to draw from it
- First, consider just L_0 norm of a single vector and reduce the problem to a nearby norm…

Zeroing in on L_0 Norm

Suppose absolute value of any entry in the vector $< B$

$$\|a\|_0 = \Sigma\ |a_i|^0 \leq \Sigma\ |a_i|^p \leq \Sigma\ B^p\ |a_i|^0 \leq B^p\ \|a\|_0$$

Setting $B^p = (1+\varepsilon)$ means

$$\|a\|_0 \leq \|a\|_p^p \leq (1+\varepsilon)\ \|a\|_0$$

So setting $p = \varepsilon / \log B$, allows approximation of L_0 by L_p – reducing p zeros in on L_0

Stable Distributions

Let X be a random variable distributed with a *stable distribution.* Stable distributions have property that

$$a_1X_1 + a_2X_2 + \ldots a_nX_n \overset{\text{in dbn.}}{=} \|(a_1, a_2, \ldots, a_n)\|_p X$$

if $X_1 \ldots X_n$ are stable with stability parameter p

Gaussian distribution is stable with parameter 2

Stable distributions exist and can be simulated for all parameters $0 < p \leq 2$.

Maintaining the Sketch

Let $x_{1,1} \ldots x_{k,n}$ be values drawn independently from stable distribution with parameter p

Sketch s is $s_1 \ldots s_k$ for small k, initially 0

Maintain $s_l = a \bullet x_l = a_1 x_{l,1} + a_2 x_{l,2} + \ldots a_n x_{l,n}$

When (i,j) in the stream arrives, update sketch:

$$s_l \leftarrow s_l + j^* x_{l,i} \text{ for } l = 1 \text{ to } k$$

Finding the L_0 Norm

Use the sketch to approximate the L_0 norm

Compute $\|\hat{a}\|_0 = \text{median}\{|s_1|^p \ldots |s_k|^p\}$

We know each s_l is distributed as $\|a\|_p X$

median $|s_l|^p$ distributed as median$(\|a\|_p^p |X|^p)$

$= \|a\|_p^p \text{ median}(|X|^p)$

Set $k = 3/\varepsilon^2 \log 1/\delta$ repeats, and bound probability that sampled values are far from true median

Probability Calculation

Let min be defined by $\Pr[|X|^p < \min] = \frac{1}{2} - \varepsilon$

If $\Pr[|X|^p < \text{median}|s_l|^p/\|a\|_p^p] < \frac{1}{2} - \varepsilon$ then median $|s_l|^p/\|a\|_p^p <$ min and at least $k/2$ values $<$ min [this is a bad event]

Set indicators $Y_l = 0$ if $|s_l|^p/\|a\|_p^p <$ min, else 1

Y_l are independent, $\Pr[Y_i = 1] = \frac{1}{2} + \varepsilon$ and $E(\Sigma Y_l) = k(\frac{1}{2} + \varepsilon)$

Bad event is indicated by $\Sigma Y_l < k/2$

Chernoff Bound

Apply the Chernoff bound here: want to know

$\Pr[\Sigma Y_l < k/2]$

$k/2 = k/2(\frac{1}{2} + \varepsilon)/(\frac{1}{2} + \varepsilon) = E(\Sigma Y_l)/(1 + 2\varepsilon)$

$\cong (1 - 2\varepsilon)E(\Sigma Y_l)$

So $\Pr[\Sigma Y_l < k/2] < \exp(-k(\frac{1}{2} + \varepsilon)\varepsilon^2/2)$

$= \exp(-3\log 1/\delta\ (\frac{1}{2} + \varepsilon)/2)$

$< \exp(-\frac{3}{4} \log 1/\delta) < \delta/2$

Using the bound

So $\Pr[\Pr[|X|^p < \text{median}|s_l|^p] < \frac{1}{2} - \varepsilon] < \delta/2$

Similar argument to show same $\delta/2$ bound for $(1+\varepsilon)$

Writing F(x) for cumulative dbn function of $|X|^p$:

$$\Pr[F(\text{median}|s_j|^p) \in [\tfrac{1}{2} - \varepsilon, \tfrac{1}{2} + \varepsilon]] > 1-\delta$$

$$\Pr[\text{median}|s_j|^p \in [F^{-1}(\tfrac{1}{2} - \varepsilon), F^{-1}(\tfrac{1}{2} + \varepsilon)]] > 1-\delta$$

$$\Pr[\text{median}|s_j| \in [F^{-1}(\tfrac{1}{2})(1-O(\varepsilon)), F^{-1}(\tfrac{1}{2})(1+O(\varepsilon))]] > 1-\delta$$

if the derivative of F is bounded around the median

Consequences

$$\Pr[\ (1-\varepsilon)\ \text{median}|X|^p \le \text{median}|s_l|^p / \|a\|_p^p \le (1+\varepsilon)\ \text{median}|X|^p] > 1-\delta$$

Overall probability we are within $(1 \pm \varepsilon)$ is $\ge 1-\delta$

Sets $k = O(1/\varepsilon^2 \log 1/\delta)$ repetitions

But… need to generate values from stable dbns

need to store all $x_{l,i} = O(kn)$ storage

Generating Stable Distributions

Compute r, θ as uniform random variables in range [0...1], [-π,π]

Chambers, Mallows, Stuck 76:

$$stable(\theta, r, p) = \frac{\sin p\theta}{\cos^{1/p}\theta}\left(\frac{\cos(\theta(1-p))}{-\ln r}\right)^{\frac{1-p}{p}}$$

stable(θ, r, p) is distributed with stable distribution with parameter p

Reducing space needs

- $x_{l,i}$ must be from stable distribution with parameter p
- $x_{l,i}$ must be the same every time it is used

Generate values from a stable distribution using the transform from uniform distributions

Use appropriately chosen pseudo-random number generator to generate r, θ as function of i,l and seed

Argue that this generates sufficient randomness, and also only limited precision is needed.

Guaranteed Accuracy

Final Result, with scaling of ε

$$(1-\varepsilon)\ \|a\|_0 \leq \text{median}(|s_l|^p)/\text{median}\ |X|^p \leq (1+\varepsilon)\ \|a\|_0$$
with probability $1-\delta$

Space usage is small: the L_0 sketch consists of $O(1/\varepsilon^2 \log 1/\delta)$ counters

Time per item is to update each counter, $O(1/\varepsilon^2 \log 1/\delta)$

Complete Algorithm

```
initialize sk[1…k] = 0.0
for all tuples (i,j) do
  initialize random with i
  for s = 1 to k do
    r1 = random(); r2 = random()
    sk[s] = sk[s]+j*stable(r1,r2,p)

for s = 1 to k do
  sk[s] = absolute(sk[s])^p
return median(sk)*scalefactor(p)
```

Simple to implement, can run quickly, small space

Properties

By linearity of the construction, all other variations are straightforward to compute:

$$sk(\mathbf{a} + \mathbf{b}) = sk(\mathbf{a}) + sk(\mathbf{b})$$
$$sk(\mathbf{a} - \mathbf{b}) = sk(\mathbf{a}) - sk(\mathbf{b})$$

by linearity of dot product, so can approximate $\|a - b\|_0$ and $\|a + b\|_0$ etc. with the same accuracy (ie $1 \pm \varepsilon$ with probability $1-\delta$).

Practical Issues: Speed

- Speed bottleneck is in generating values from stable distributions
- Use results from statistics to generate values from a simpler dbn which have the same distribution
- For any distribution, sum of multiple copies of the same distribution will always tend to be a stable distribution
- So look for a dbn that is in the "domain of attraction" of a stable dbn with parameter p

Optimization

- Extended central limit theorem: a dbn is in the domain of attraction of a stable dbn with parameter p if F, its CDF, obeys

$$1 - F(x) + F(-x) = x^{-p}\, O(1)$$

- Let U = Uniform(-1,1), and set $Y = \text{sign}(U) * |U|^{-1/p}$
- Then Y is in the domain of attraction of X(p).
- Replace X(p) with Y in the algorithm to speed up

Practical Observations

- Accuracy is pretty good in practice
- Outperforms Flajolet-Martin probabilistic counting to find distinct elements in accuracy for same space usage
- With faster generation of values, also competitive in time
- Other distinct element methods are asymptotically faster, but less flexible

Applications

- Now consider an application of L_0 techniques to solve a novel problem.
- Consider data streams which consist of many signals
- Reduce the problem to several L_0 norm computations in small space

Multiple Signals

Previous work considers only a single signal at a time

Many data streams consist of multiple signals from several distributions, from which we want to extract some global information

Examples:

- financial transactions from many different individuals
- web clickstreams from many users registered on different machines
- multiple readings from multiple sensors in atmospheric monitoring

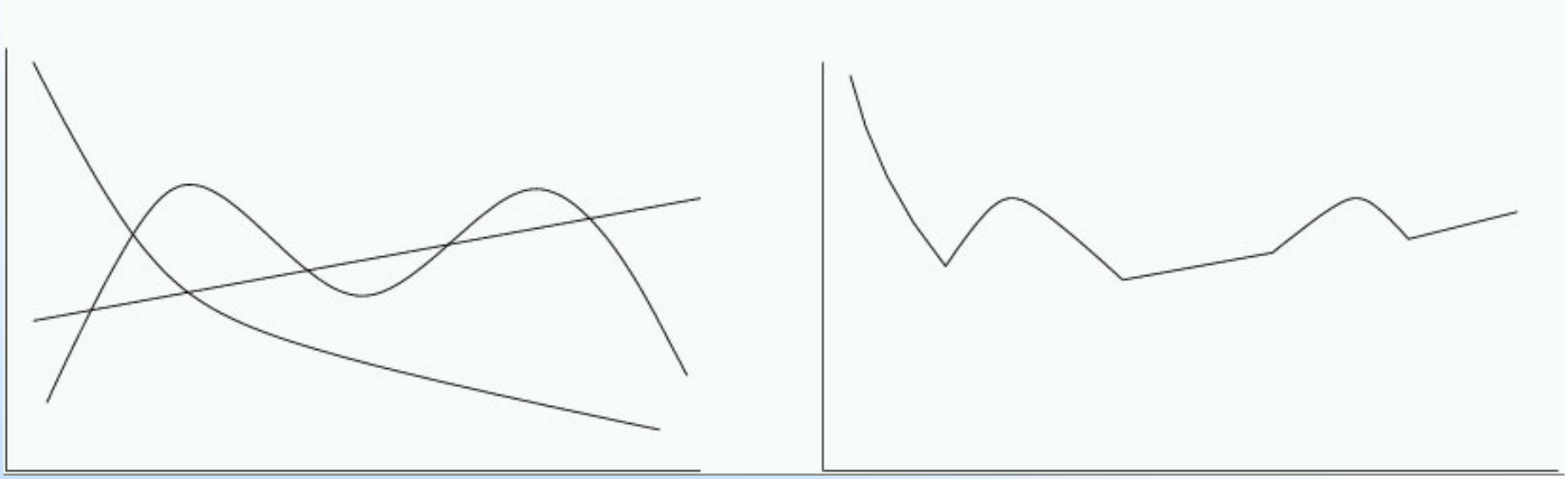
Multiple Signal Model

- Model stream of multiple signals S as simply structured series of items
- n items in the stream S=(i, a[i,j]) means a[i,j] is the value of distribution j at location i
- Assume: a[i,j] is bounded by polynomial in n
- Don't assume that j is made explicit in stream or that we see updates for every [i,j] pair.
- Number of signals and domain of signals is too large to store explicitly

Dominance Norm

- The dominance norm measures the "worst case influence" of the different signals
- Defined as $\mathrm{Dom}(S) = \Sigma_i \max_j \{a[i,j]\}$
- A function of the marginals of a matrix of the signal values
- Can also think of this as the L_1 norm of the upper-envelope of the signals

Dominance Norm



- Maximum possible utilization of a resource
- Applied in financial applications, electrical grid
- Treat as an indicator of actionable events

Dominance Norm

- Suppose each a[i,j] is 0 or 1
- Consider each signal to define a set X_j, then

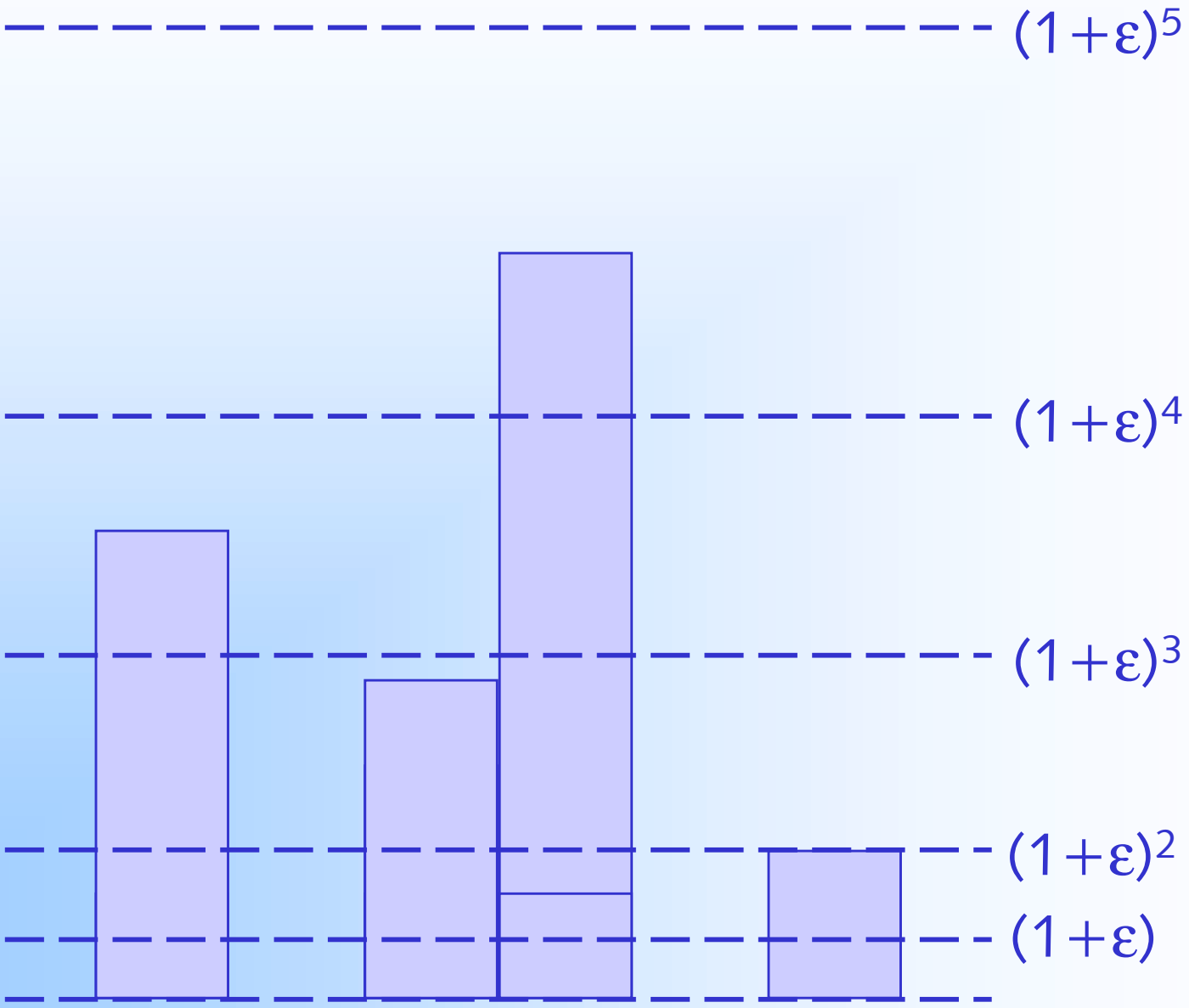
$$\text{Dom}(S) = |\cup_j X_j|$$

This can be solved with stream algorithms for finding unions of multiple sets

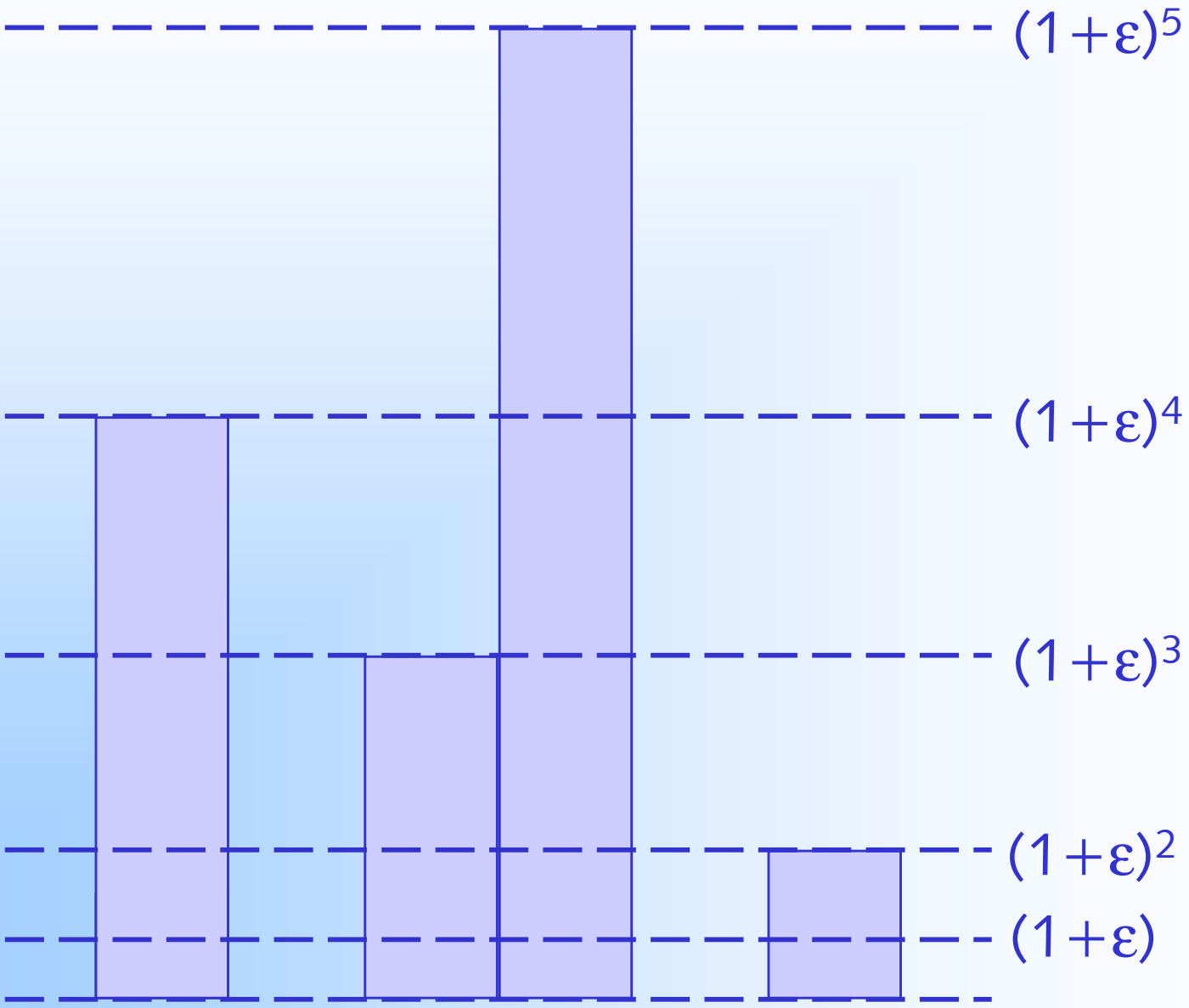
Can also be thought of as counting the number of distinct items i in the stream

Can this be generalized for arbitrary a[i,j]?

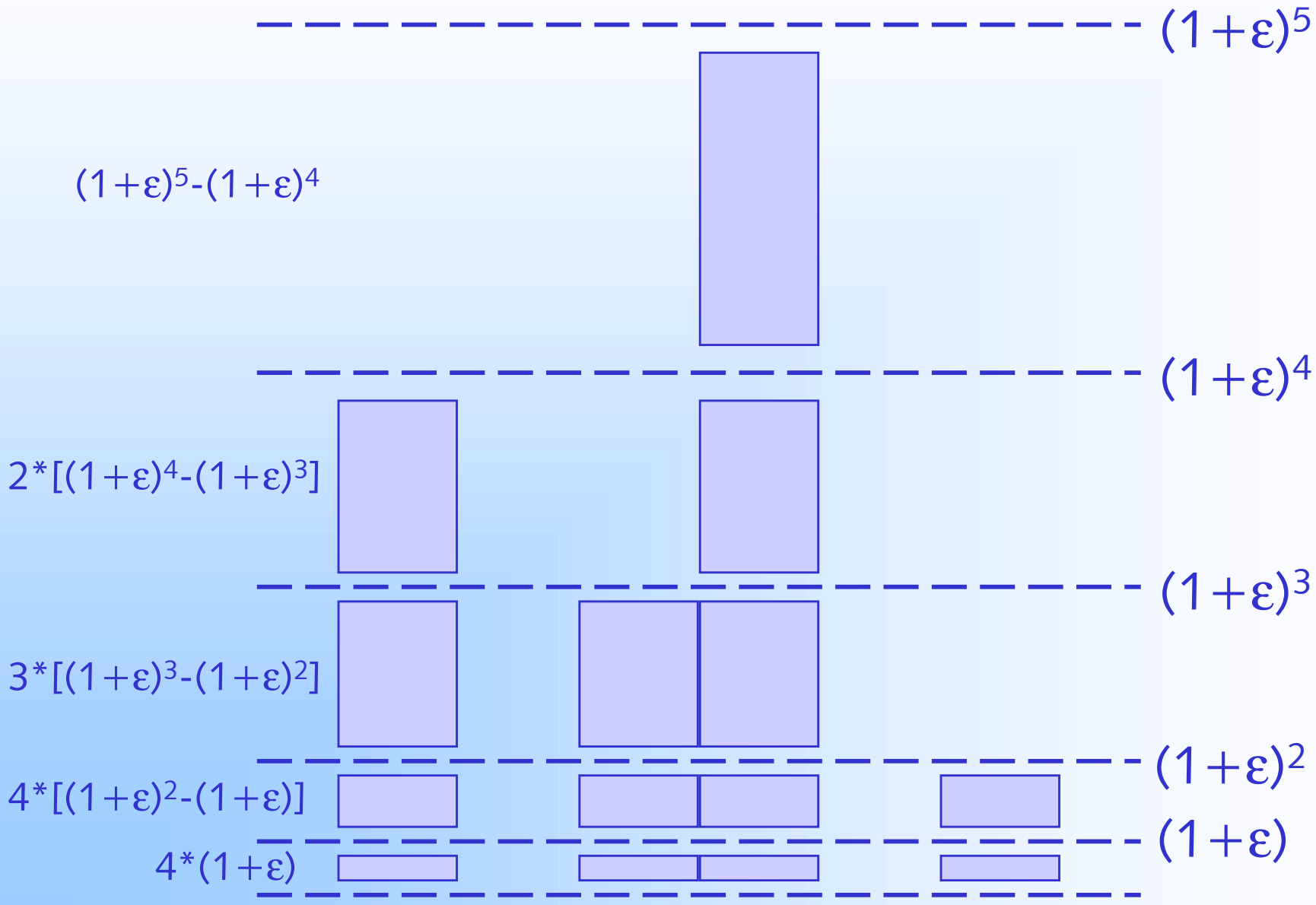
Approximation



Approximation



Approximation



Space Cost

- $\log_{1+\varepsilon}$ (max val / min val) distinct element algorithm instances = O(log (n)/ε) instances
- Space required is O(poly-log(n)/ε^2) per instance using existing techniques
- Total space is O(poly-log(n)/ε^3)
- Cubic space dependency on 1/ε is high – can do better by modifying L_0 norm solution

Approximation Algorithm

- Make new sketch, set $s_l = 0$ initially
- Let $x_{i,k}$ be values drawn from Stable Distribution with parameter $p = \varepsilon/\log n$.
- For every (i,a[i,j]) in the stream,

$$z = z + \Sigma_{k=1}^{a[i,j]} x_{i,k}$$

- Repeat independently in parallel $O(1/\varepsilon^2 \log 1/\delta)$ times, then with probability at least $1-\delta$,

$$(1-\varepsilon)\mathrm{Dom}(S) \leq \frac{\mathrm{median}(|z|^p)}{\mathrm{median}(|X|^p)} \leq (1+\varepsilon)\mathrm{Dom}(S)$$

Efficient Computation

- Direct implementation means adding a[i,j] values to the counters for every update
- But, each value is drawn from a stable distribution, and we know sum of stables is a stable
- Use similar trick to before, round to nearest power of $(1+\varepsilon)$ and add the $O(\log(n)/\varepsilon)$ values to the counters
- So update time is $O(\log(n)/\varepsilon^3)$

Full results

- Approximate the Dominance norm within $1 \pm \varepsilon$ with probability at least $1-\delta$ using $O(1/\varepsilon^2 \log (1/\delta))$ counters
- Time per update is $O(1/\varepsilon^3 \log (1/\delta))$
- Possible to 'subtract off' the effect of earlier insertions – not possible with most distinct element algorithms

Other Dominances

- Natural questions: are other notions of dominance on multiple streams tractable?
- Take Min-Dominance:

$$\text{MinDom}(S) = \Sigma_i \min_j \{a[i,j]\}$$

- Let X_1, X_2 be subsets of $\{1...n/2\}$. Set $a[i,j]=1 \Leftrightarrow i \in X_j$
- Then $\text{MinDom}(S) = |X_1 \cap X_2|$
- Requires $\Omega(n)$ space to approximate, even allowing probability, several passes etc.

Extensions

- Other reasonable definitions of dominances – eg Median Dominance, Relative Dominance between two streams, also require linear space
- Are there other natural quantities which are computable over streams of multiple signals?
- What quantities are good indicators for actionable events?

Closed Problems?

- $\Omega(1/\varepsilon^2)$ space is necessary [see David Woodruff's talk today] but what about time?
- $O(1/\text{poly}(\varepsilon))$ time is OK for database speeds, but too long for network speeds.
- New technique from Piotr Indyk improves update time to $O(1)$
- Method is more related to that of Flajolet-Martin, with some clever use of hashing

Open Problems

- To find other applications for L_0 and related metrics – L_p generally not much studied for non-integer p
- To find new applications and improvements of stable distributions to metric spaces and embedding problems
- To find more uses of statistical distributions in this area – approximate other normed spaces with similar techniques
- To find some lunch