Stable Distributions for Stream Computation:

It's as easy as 0,1,2

Graham Cormode

graham@dimacs.rutgers.edu

dimacs.rutgers.edu/~graham

"I had a really good night last night. I found I can count up to 1023 on my fingers."

Chris Hughes, http://www.jacobite.org.uk/dave/odd/chrisism.html

Stable Distributions

Stable distributions have the (defining) property

$$a_1X_1 + a_2X_2 + a_3X_3 + \ldots a_nX_n$$

is distributed as $\|(a_1, a_2, a_3, \ldots, a_n)\|_pX$

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 $0 < p < 2$.

"A physicist would fly across the Atlantic in one hour but might fall out of the sky. A mathematician would fly across the Atlantic in ten hours but would be sure he wouldn't fall out of the sky."

Stable Sketches

Using stable distributions, can make sketches of vectors [Indyk00]

Sketch of vector $\mathbf{a}$ = sk($\mathbf{a}$), sketch has dimension $O(1/\varepsilon^2 \log 1/\delta)$

Main property:

Use sk($\mathbf{a}$) to find a_p such that, with prob $1-\delta$

$$(1-\varepsilon)\|\mathbf{a}\|_p \leq a_p \leq (1+\varepsilon)\|\mathbf{a}\|_p$$

"With probability zero I am a penguin. I mean, you don't get that normally."

"We want to count in units of twelve, so we need an extra digit on each finger"

Sketch Properties

- Compute sketch from implicit stream representation of **a**, as sequence of updates
- Linear transform, so sketches can be combined linearly:

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

- Good in practice [C,Indyk, Koudas, Muthukrishnan, 02], some code on my webpage

"This is a genuinely true story told to me by somebody. I don't know whether I believe him."

Sketch Applications

Sketches have many direct applications:

- Efficient communication of diagnostics on networks
- Small space representation of massive data (a kind of dimensionality reduction)
- Speed up data mining etc. – instead of clustering with large vectors, cluster with small sketches

"If we could link the computers, we could play solitaire against each other... I was thinking of a competitive game, but I couldn't think of one."

Pause for thought

Half way through, so time for a quick break…

…did you enjoy it?

Remainder of the talk: further applications, based on choice of parameter p

- 0: Distinct Elements
- 1: Embeddings
- 1-2: Wavelets
- 2: Nearest Neighbor

"English is an illogical language, because we can have two statements meaning different things."

0: L_0 Norm for Distinct Elements

What happens as parameter p tends to 0?

- $(\|\mathbf{a}\|_p)^p = \Sigma\ |a_i|^p = 1$ if a_i is nonzero, else 0
- So we can count the number of nonzero entries in x, as items arrive and depart
- Flexible way to track distinct items in a stream
- What is $\|\mathbf{a} - \mathbf{b}\|_0$? Counts number of places **a** and **b** differ: "Hamming Norm": useful measure of similarity

[C,Datar,Indyk,Muthukrishnan,02/03]

"And the knee-bone's connected to the wrist bone."

1: L_1 for Embeddings

- With $p=1$, sketches are like a dimensionality reduction for L_1
- Build approximation algorithms for many metric spaces using this pattern: embed items into (high-dimensional, sparse) L_1, reduce to low-dimension using sketches
- Eg. Approximate clustering, nearest neighbors on some string and permutation edit distances, sketches computable in stream

[C,Muthukrishnan'02;C,Muthukrishnan,Sahinalp'01]

"I've worked out trousers. It's a double-ring doughnut, of course!... So pants aren't so funny any more."

1-2: Wavelets on Streams

- Compute a good (Haar) B-term wavelet representation of a massive streaming vector, using sketches
- Requires computing a sketch of [0…01…10…0], can be done efficiently with range-summable stable variables
- Follows from the fact that sum of stables is stable, and drawing values conditioned on their sum

[Gilbert, Guha, Indyk, Kotidis, Muthukrishnan ’02]

"What do you call a coat that you go out in in the rain in when it's not an umbrella?"

2: Approx Nearest Neighbors

- Can use stable distributions to construct "locality sensitive hash functions"
- These plug into approximate nearest neighbors scheme of Indyk-Motwani
- Whole thing can be computed on the stream: for database and query points, compute hashs, store / query stored hashs.

[Datar, Immorlica, Indyk, Mirrokni, '02]

"It must be yesterday, I mean Friday. That's thinking of yesterday as last working day and ignoring the weekend."

Extensions and Open Problems

- Are there other distributions that have similar properties for other functions – eg "log stable": distributed as $\Sigma \log (a_i) X$?
- Faster, more numerically stable simulation of stable distributions for non-integer p (some progress for $p \rightarrow 0$)
- Range summability for all p? (some results for sums from $0 \ldots k$)

"It's only cryptic if you don't know what it means."