Stable Distributions for Stream Computations: it's as easy as 0,1,2

Graham Cormode*

-1. Introduction

A surprising number of data stream problems are solved by methods involving computations with stable distributions. This paper will give a short summary of some of these problems, and how the best known solutions depend on use of stable distributions; it also lists some related open problems. Stable distributions arise from seeking statistics with the property that $\sum_{i=1}^{n} a_i X_i$ is distributed as $(\sum_{i=1}^{n} |a_i|^p)^{1/p} X_0$. Here, a_i are scalars, and $X_0 \ldots X_n$ are independent and identically distributed random variables. Such distributions[1] exist for all $p \in (0, 2]$. Gaussian is stable with $p = 2$ and Cauchy stable with $p = 1$. See the books [10, 11] for a statistical treatment of Stable Distributions. The principal application to a stream context is to approximate of the L_p norm of a stream of values defining a vector. That is, the computation of $||a||_p = (\sum_i |a_i|^p)^{1/p}$, where a may be described in some arbitrary, incremental manner. This approach was pioneered by Indyk[2].

Theorem 1 (Theorem 1 of [8]). *Given $0 < p \leq 2$ and a sequence of updates to an (initially zero) vector a of the form (i, d_j), which we interpret as "add d_j to entry a_i", we can compute a small "sketch" of the vector a, $sk(a)$. The sketch is a vector with $O(1/\epsilon^2 \log 1/\delta)$ entries. With it, we can compute an approximation of $||a||_p$ which is correct within a factor of $(1 \pm \epsilon)$ with probability $1 - \delta$.*

The sketch is computed by forming the dot-product of the vector a with a matrix r, where each entry of r is drawn independently from a stable distribution with parameter p. Each entry in the sketch is taken absolutely, so is distributed as $||a||_p |X|$. The computation of $sk(a) = a \cdot r$ from a stream of updates to a relies on a number of technical issues, including

- That the dot product is a linear transformation, and so updates require only a scalar multiple of a row of r to be added to sk.
- That stable distributions can be simulated for any value of p using transforms from uniform variables [1]. [3]
- That it suffices to use pseudorandom generators with small space to generate row i as a function of i.
- That the median of $O(1/\epsilon^2 \log 1/\delta)$ estimators is close to the median of the distribution with probability at least $1 - \delta$, relying on the derivative of the distribution being bounded at the median.

These results are shown for $p = 1$ and $p = 2$ in [8], and for all $0 < p \leq 2$ in [4]. These give an efficient way to approximate the L_p norm of a datastream with small space requirements. Experimental work has shown this to be accurate in practice [4]. The method has many attractive features, in particular that because of the linearity of the method of construction, sketches can be combined by summing them component wise to find the aggregation over multiple streams, and more importantly, taking differences component-wise allows the approximation of L_p differences, $||a - b||_p$. This follows since $sk(a) + sk(b) = sk(a + b)$, and $sk(a) - sk(b) = sk(a - b)$. This leads to efficient distributed *communication* schemes for these problems, since different parties can communicate their sketches with a cost linear in the size of the sketch.

These results are of interest in themselves, but we go on to describe how they have been applied to several other problems. The power and flexibility of these distributions has meant that they have found numerous applications, and they have shown impressive performance improvements and additional functionality when compared to existing solutions. Conversely, in the course of their application to streaming questions, new results have been proved about stable distributions (such as range-summable constructions), which contribute back to the statistical community. Hercules showed how to process stables with streams in just one day[4]; a less herculean task is to show how to process streams with stable(distribution)s in two pages. [9] is a longer survey of data stream problems, with additional background to the following discussion. In Sections 0, 1 and 2 we discuss results specific to stable distributions with $p = 0, 1$ and 2, respectively. Note that *fractional* values of p have been investigated for data mining purposes [4].

0. L_0 for Distinct Elements

It is straightforward to observe an interesting behaviour of the approximation of $||a||_p$ raised to the power p as $p \to 0$: if a_i is zero, then this contributes nothing; if a_i is non-zero, then the contribution of $|a_i|^p$ is close to 1. Then, $||a||_p^p$ approximates the number of non-zero entries in the vector a. If we add 1 to a_i whenever some item labelled i arrives in the stream, and subtract 1 whenever item i departs, then the number of non-zero entries in a is precisely the current number of distinct elements: a fundamental quantity required in database management and network scenarios. This approach was described in [3]. Additionally, a faster and more robust way of generating values from stable distributions was tested, based on the limiting distribution as $p \to 0$ being $[\text{Uniform}(-1, 1)]^{-1/p}$. This

*graham@dimacs.rutgers.edu

[1] Technically, we are describing *symmetric and strictly stable* distributions.

[2] On being described as a pioneer, Piotr commented "I can imagine myself in a transatlantic ship coming to Boston, in a funny hat, holding a bible and looking forward to the New World".

[3] The author's implementation of this is available from `http://athos.rutgers.edu/~muthu/stream-seminar.html`

[4] Hercules cleaned the Augean stables by diverting a river through them. See, for example, *Hercules* (Disney, 1997).

idea was extended in [5] in order to compute the worst case influence of multiple data streams, defined as $\sum_i \max_j(a_{i,j})$. [5] began a study of the behavior of algorithmic applications of stable distributions as $p \to 0$, and computes certain range sums of stable distributions in a simple fashion.

1. L_1 for Embeddings

Theorem 1 can be thought of as a dimensionality-reduction for vectors, an analog of the Johnson-Lindenstrauss lemma for other L_p norms[5]. Because of the highly flexible way in which the sketch can be updated, it can be used in many situations as a "black box" for stream computations to reduce the size required when intermediate results can be modelled as vectors in L_1 space. [2] gives several examples where a single pass over a string computes a vector representation so that an edit distance between strings is approximated by the L_1 distance between vectors. The embedding is generated as a series of additions to the vector; by sketching, the space required is reduced from the exponential size of the vector to the effectively constant size of the sketch.

Range-summability of stable distributions for the Cauchy ($p = 1$) and Gaussian ($p = 2$) cases is shown in [7]. This is used to help find Haar wavelet coefficients: the sketch corresponding to a wavelet vector of the form $\pm(0^j 1^k 0^{n-(j+k)})$ can be found in time $O(\log n)$ instead of $O(k)$ by computing the range sum $\sum_{i=j+1}^{k} X_i$. This relies on the defining property of stable distributions, that the sum of stable distributions is itself distributed stable, and so constructing pairs of random variables by first drawing their sum from an appropriate distribution, and then picking the pair conditioned on this sum.

2. L_2 for Nearest Neighbors

Stable distributions have found recent application in improving results on Approximate Nearest Neighbor searching [6]. This works by making sketches, and then "coarsening" each entry onto an integer range, giving a hash function whose probability of collision is related to the similarity of the vectors. This then feeds into the Locality-Sensitive Hashing method of Indyk (see [6]). Since it depends on sketches, then all the necessary computations can be made on the stream, so for streams of input vectors arriving in arbitrarily interleaved order, we can find approximate nearest neighbors for each on the fly.

π. Open Problems

There still remain some important questions to resolve to make a complete theory of stable distributions used as tools in algorithmic processing.

- Strong range-summability results are known for L_1 and L_2; can these be extended to all $0 < p \leq 2$?[6]
- Existing work makes numerical approximations of the median of $|X|$ and assumes that the derivative at this point is bounded for non-integer p; an analytical approach would be preferable, especially a characterization of the behavior of the median as p tends to 0.[7]
- Are there other distributions which are comparable to stable distributions which would allow computation of other quantites of interest? For example, might there be log-stable distributions where $\sum_i a_i X_i$ is distributed as $(\sum_i \log a_i) X_0$ or $(\sum_i a_i \log a_i) X_0$? [8] Even if these do not exist as distributions, can we build constructions making use of L_p norm estimations as building blocks?
- Stable distributions are known not to exist with parameter $p > 2$, and strong space lower bounds are known for computing $||a||_p$ on streams for (integer) $p \geq 3$. Can lower bounds tell us more about distributions with certain properties, and vice-versa?
- Computations of values from stable distributions can be slow and numerically unstable, since the formula in [1] is somewhat complex. Can generation of values be made faster using implementation tricks, look-up tables, limited precision, or combinations of these?[9]

Lastly, it will be of interest to deploy computations using sketches from stable distributions in "real world" scenarios, in software or dedicated hardware.

References

[1] J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.

[2] G. Cormode. *Sequence Distance Embeddings*. PhD thesis, University of Warwick, 2003.

[3] G. Cormode, M. Datar, P. Indyk, and S. Muthukrishnan. Comparing data streams using Hamming norms. *IEEE TKDE, 2003*.

[4] G. Cormode, P. Indyk, N. Koudas, and S. Muthukrishnan. Fast mining of tabular data via approximate distance computations. In *Proceedings of ICDE*, pages 605–616, 2002.

[5] G. Cormode and S. Muthukrishnan. Estimating dominance norms of multiple data streams. DIMACS Tech Report 2002-35.

[6] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. Manuscript, 2002.

[7] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings of 34th ACM STOC*, 2002.

[8] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proceedings of the 40th IEEE FOCS*, pages 189–197, 2000.

[9] S. Muthukrishnan. Data streams: Algorithms and applications. Invited talk at *14th ACM-SIAM SODA*, 2003 Available from `http://athos.rutgers.edu/~muthu/`.

[10] V. V. Uchaikin and V. M. Zolotarev. *Chance and Stability: Stable Distributions and their applications*. VSP, 1999.

[11] V. M. Zolotarev. *One Dimensional Stable Distributions*, volume 65 of *Translations of Mathematical Monographs*. American Mathematical Society, 1983.

[5] It is not a perfect analog, since the use of the median operation to extract the result means that it is not an embedding into a normed space; for many practical applications, this is not a significant disadvantage.

[6] The range summability shown in [5] applies only to sums from zero.

[7] Since we know that in the limit the median is ln 2 [5].

[8] Sum of logs would be of interest for computing relative changes, and $a_i \log a_i$ for computing the empirical entropy of streams

[9] See [3] for some improved generation results on the $p \to 0$ case.