HIGH FREQUENCY MOMENTS VIA MAX-STABILITY

Alexandr Andoni

Columbia University

ABSTRACT

We present a new, simple algorithm for sketching the $k > 2$ frequency moment of a dynamic stream, or simply the ℓ_k norm of a vector in the linear sketching model.

The new algorithms are based on exponentially distributed random variables, which possess a certain "max-stability" property, similar in spirit to the "p-stability" property used in [Indyk, JACM'06] for sketching ℓ_k norms for $k \leq 2$.

Our resulting sketching algorithm can be seen as a "weak embedding" of an n-dimensional ℓ_k space into ℓ_∞ space of dimension $m = O(n^{1-2/k} \log n)$: it preserves the norm of a vector up to constant approximation, with constant probability. We note that this dimension is optimal for linear embeddings (sketches) with constant approximation, as shown in [Andoni-Nguyen-Polyanskiy-Wu, ICALP'13].

The preliminary version of this result has appeared as a blog post in 2012, and its main idea has since been used in other streaming algorithms.

***Index Terms*—** sketching, dimension reduction, metric embeddings, streaming

1. INTRODUCTION

An important notion in modern algorithmic design is that of *sketching*. Sketching is a method for summarizing complex objects into smaller ones so that the summaries still capture properties relevant to the particular algorithmic task at hand. For example, the prototypical use of sketching is for summarizing high-dimensional vectors into small summaries that are nonetheless useful for *distance estimation*. The classic such example is the Johnson-Lindenstrauss Lemma for dimension reduction [1], which shows that high-dimensional vectors can be summarized (sketched) into vectors of smaller dimension, proportional to the log of the set size. Sketching has since found many applications, for example for the streaming model [2], nearest neighbor search, or compressive sensing, where acquisition of a signal can be viewed as a (linear) sketching algorithm. More recently, sketching has been used to speed up computational tasks in areas such as numerical linear algebra [3], and dynamic graph algorithms [4, 5].

Most of the aforementioned applications rely on perhaps the most classic setting for sketching: sketching the ℓ_k norm of a vector $x \in \mathbb{R}^n$. This problems dates back at least to the seminal work of [6] who provided the first such results. The formal definition of linear sketching for ℓ_k norms is as follows:

Definition 1.1. *Fix $k \geq 1$ and approximation $1 + \epsilon$, as well as dimension $n \geq 1$ and sketch complexity $m \geq 1$. We say there is* a sketch *for ℓ_k, if there exists a distribution over linear sketching functions $f : \mathbb{R}^n \to \mathbb{R}^m$, and a reconstruction procedure $R : \mathbb{R}^m \to \mathbb{R}$ such that, for any $x \in \mathbb{R}^n$, we have that*

$$\Pr[R(f(x)) = (1 \pm \epsilon)\|x\|_k] \geq 2/3.$$

We call m to be the sketching complexity of the sketch f. (Often times, the reconstruction procedure R is deterministic.)

Note that such a linear sketch can be used for estimating distances between two vectors $x, y \in \mathbb{R}^n$ *from their sketches only* due to the following identity: $R(f(x) - f(y)) = R(f(x - y))$, which, by definition, approximates $\|x - y\|_k$.

It is hence natural that the sketching complexity of ℓ_k has been heavily studied. When $k \in (0, 2]$, the exact best sketching complexity is now well-understood; see [7], the references therein, as well as the book [3].

In this paper we focus on the case of $k > 2$, for which the exact bound (with respect to both the dimension n and approximation $1 + \epsilon$) is still open, despite significant efforts. The first sublinear-space algorithm in this regime, due to [6], gave a space bound $n^{1-1/k} \cdot (\log n)^{O(1)}$, and further showed the first polynomial lower bound for k sufficiently large. A lower bound of $\Omega(n^{1-2/k})$ was shown in [8, 9], and it was (nearly) matched in [10], who gave an algorithm using space $n^{1-2/k} \cdot (\log n)^{O(1)}$. Further research reduced the space bound to essentially $O(n^{1-2/k} \cdot \log^2 n)$ [11, 12].

For the regime of constant ϵ, the tight bound of $m = O(n^{1-2/k} \log n)$ was finally shown in [13] (a nearly-tight bound was also independently shown in [14]). A matching lower bounds was proven in [15]. Later [16] improved the dependence on ϵ for sub-constant ϵ, obtaining $m = O(n^{1-2/k} \cdot (\epsilon^{-2} + \epsilon^{-4/k} \log n))$. This is tight for $\epsilon < 1/(\log n)^{O(1)}$, matching the lower bound from [17]. See also related work in [12], [18], and [19] (where the authors manage to obtain a smaller space complexity for insertion-only streams, which is a somewhat different setting than considered here).

Main result. Here we present a new, simple algorithm for linearly sketching the ℓ_k norm for $k > 2$. The algorithm matches the bounds of $m = O(n^{1-2/k} \log n)$ for constant ϵ from [13, 16], and its main advantage is its simplicity. In fact, since the announcement of this result (in a blog post in 2012 [20]), the main idea from here has been used in other contexts [3, 21]. The main theorem follows.

Theorem 1.2. *Let $n \geq 1$, and $k > 2$ be a constant. For $m = cn^{1-2/k} \log n$ for a large enough constant c, there exists a randomized linear map $f : \mathbb{R}^n \to \mathbb{R}^m$ such that, for any $x \in \mathbb{R}^n$, we have that the image $\|f(x)\|_\infty = \max_i |f(x)_i|$ is a constant approximation to $\|x\|_k = (\sum_i |x_i|^k)^{1/k}$ with probability at least 2/3.*

Note that the sketch from above is in fact a "weak embedding" of ℓ_k into (lower dimensional) ℓ_∞: it is a linear map into ℓ_∞, which preserves the norm (up to a constant factor), with constant probability. We remark that the approximation can be improved to $1 + \epsilon$.

From a technical perspective, the algorithm relies on the approach from [13] (itself based on the ideas from [10]), and [22]. The main idea is to use the *max-stability* of exponentially distributed random variables. This can be seen as the counterpart of the p-stability notion introduced in [23] to give the first sketching algorithms for ℓ_k norms for $k \in (0, 2)$. See also Section 5 for a further discussion.

Unlike in [13], the algorithm from here bypasses the precision sampling lemma, although the latter can also be simplified using the max-stability concept; see Section 4.

2. SKETCHING ALGORITHM AND ANALYSIS

We start by presenting the sketching algorithm for the ℓ_k norm, and then proceed to analyze it, thus proving Theorem 1.2.

2.1. Algorithm

We construct the function $f : \mathbb{R}^n \to \mathbb{R}^m$ as follows. Let $x \in \mathbb{R}^n$ be the input vector. The algorithm just multiplies x entry-wise by some scalars, and then folds the vector into a smaller dimension m using standard hashing. Formally, in step one, we compute $y \in \mathbb{R}^n$ as

$$y_i = x_i / u_i^{1/k}$$

where random variable u_i is drawn from an exponential distribution e^{-u}. In step two, we compute $z \in \mathbb{R}^m$ from y using a random hash function $h : [n] \to [m]$ as follows:

$$z_j = \sum_{i:h(i)=j} \sigma_i \cdot y_i$$

where σ_i are just random ± 1. From now on, *bucket* j denotes elements i with $h(i) = j$.

The output is $f(x) = z$. In matrix notation, $f(x) = PDx$, where D is a diagonal matrix with entries $\sigma_i / u_i^{1/k}$ and P is a sparse $0/1$ "projection" matrix describing the hash function h.

2.2. Analysis

The analysis proceeds in two steps. We first show that the infinity norm of y is correct, namely that $\|y\|_\infty \approx \|x\|_k$, and then that the infinity norm of z is correct as well, i.e., $\|z\|_\infty \approx \|y\|_\infty$ (both with constant approximation with constant probability of success). In other words, step one is an embedding into ℓ_∞, and step two is a dimensionality reduction in ℓ_∞.

The claim about the infinity norm of y follows from the stability property of the exponential distribution: if $u_1, \ldots u_n$ are exponentially distributed, and $\lambda_i > 0$ are scalars, then $\min\{u_1/\lambda_1, \ldots u_n/\lambda_n\}$ is distributed as u/λ where u is also an exponential and $\lambda = \sum_i \lambda_i$.

Now, applying this stability property for $\lambda_i = |x_i|^k$ we get that $\|y\|_\infty^k = \max_i |x_i|^k / u_i$ is distributed as $\|x\|_k^k / u$. Hence, $\|y\|_\infty \in [\frac{1}{2}\|x\|_k, 2\|x\|_k]$ with probability at least $e^{-1/2^k} - e^{-2^k} \geq 0.75$.

Note that we already obtain a weak embedding of ℓ_k^n into ℓ_∞^n (i.e., no dimensionality reduction). We proceed to show that the dimension-reducing projection does not hurt.

We will now analyze the max-norm of z. The main idea is that the large entries of y will go into separate buckets, while the rest of the "stuff" (small entries) will give only a minor contribution to each bucket. Hence, the biggest entry of y will stick out in z as well, and nothing bigger will stick out, approximately preserving the max-norm. For simplicity of notation, let $M = \|x\|_k$, and

note that the largest entry of y is within a factor 2 of M with probability at least 75% (as we argued above).

What is big? We say that "big" is an entry of y such that $|y_i| \geq M/(c \log n)$ (and "small" otherwise).

Claim 2.1. *Let $l \geq 1$. In expectation, there are at most l^k indices such that $|y_i| \geq M/l$.*

Proof. $\Pr[|y_i|^k \geq M^k/l^k] = \Pr[l^k \cdot |x_i|^k/M^k \geq u_i] = 1 - e^{-l^k |x_i|^k/M^k}$. Hence the expected number of big entries in y is: $\sum_i \Pr[|y_i|^k \geq M^k/l^k] \leq \sum_i l^k \cdot |x_i|^k/M^k = l^k$. □

Hence, by Markov's, there are only $O(\log n)^k$ such big entries with at least 99% probability. Furthermore, by assumption $O(\log n)^k \ll n^{1/2-1/k} < \sqrt{m}$, with $1 - o(1)$ probability, there are no collisions among all the big entries, i.e., they all go into different buckets under the hash function $h : [n] \to [m]$.

Now, let us focus on the "extra stuff", i.e., the contribution of the small entries. Let $S \subset [n]$ be the set of small entries of y, i.e., $S = \{i \mid |y_i| < M/(c \log n)\}$. Fix some bucket index $j \in [m]$. We would like to show that the contribution of entries from S that fall into bucket j is small, say, less than $M/4$ (half the max entry of y).

Let's look at $z'_j = \sum_{i \in S : h(i) = j} \sigma_i y_i$. The expectation of z'_j is zero because $\sigma \in \{\pm 1\}$. Also the variance is

$$\mathbb{E}_{h,\sigma}\left[{z'}_j^2\right] = \mathbb{E}_h\left[\sum_{i \in S : h(i) = j} y_i^2\right] \leq \|y\|_2^2/m.$$

We'd like now to relate $\|y\|_2$ to $M = \|x\|_k$. Here comes the exponential distribution at rescue again. Note that $\mathbb{E}_{\{u_i\}_i}[\|y\|_2^2] = \sum_i x_i^2 \cdot \mathbb{E}_{u_i}[1/u_i^{2/k}] = \sum_i x_i^2 \cdot O(1) = O(\|x\|_2^2)$ since the expectation $\mathbb{E}[1/u^{2/k}]$ for an exponentially distributed u is constant. Together with standard inter-norm inequality that $\|x\|_2^2 \leq n^{1-2/k}\|x\|_k^2 = n^{1-2/k} \cdot M^2$, we have that $\mathbb{E}[\|y\|_2^2] \leq O(n^{1-2/k} M^2)$. By Markov's, we have that $\|y\|_2^2 \leq O(n^{1-2/k} M^2)$ with probability 99%.

Now we can complete computing the variance of z'_j to get $\mathbb{E}[{z'}_j^2] \leq O(n^{1-2/k} M^2/m) \leq O(M^2/(c \log n))$. One can now apply Chebyshev's inequality and conclude that the "extra stuff" in a bucket j is $|z'_j| \leq o(M)$ with constant probability.

This is however not enough to complete the proof: we would need the above for *all* buckets j at the same time, and, in particular, we would like to have $|z'_j| \leq M/4$ for a fixed j with high probability (not just constant). To achieve this, we use a stronger concentration inequality, namely the Bernstein inequality, applied to the elements $i \in S$, for which $|y_i| \leq M/c \log n$.

In particular, for a fixed bucket j, we analyze the sum $z'_j = \sum_{i \in S : h(i) = j} \sigma_i y_i$, where each $\mathbb{E}[z_j] = 0$ and $\mathbb{E}[{z'}_j^2] \leq O(M^2/(c \log n))$. Then, by Bernstein's inequality, we have that, for $\alpha = 1/4$,

$$\begin{aligned} \Pr[|z'_j| > \alpha M] &\leq \exp\left[-\frac{(\alpha M)^2/2}{\mathbb{E}[{z'}_j^2] + \frac{M}{c \log n} \cdot \frac{\alpha M}{3}}\right] \\ &\leq \exp\left[-\Theta(\alpha c \log n)\right]. \end{aligned}$$

For c large enough, we obtain a high probability statement as desired.

Concluding, we have that $\|z\|_\infty \in [\|x\|_k(1/2 - 1/4), \|x\|_k(1/2 + 1/4)]$ with probability at least $0.75 - 0.01 - o(1) - 0.01 - o(1) \geq 2/3$. This completes the proof of Theorem 1.2.

3. OTHER ASPECTS OF THE SKETCH

There are two aspects of the sketch which we would like to comment on now. First, we can in fact also obtain a $1 + \epsilon$ approximation with just slightly more work. Second, we discuss the use of randomness use (which is an important aspect for streaming algorithms).

Achieving $1 + \epsilon$ approximation. We note that it is simple to obtain $1 + \epsilon$ approximation, though the resulting sketching is not a (weak) embedding into ℓ_∞ anymore. To obtain better approximation, we take $r = O(1/\epsilon^2)$ sketches f, each with target dimension $m = O(n^{1-2/k} \log n \cdot \epsilon^{-2})$. The estimation procedure just computes the value (i.e., the ℓ_∞ norm of the m coordinates) for each of the r sketches, and outputs the median value (normalized by a factor of $\ln 2$).

The main claim is that, in fact, for each sketch, we have the following two statements:

$$\Pr[\|y\|_\infty > (1 - \epsilon) \cdot \ln(2) \cdot \|x\|_k] \leq 1/2 + O(\epsilon),$$

$$\Pr[\|y\|_\infty < (1 + \epsilon) \cdot \ln(2) \cdot \|x\|_k] \geq 1/2 - O(\epsilon).$$

Furthermore, using Bernstein's inequality with $\alpha = O(\epsilon)$ and $c = \Theta(1/\epsilon^2)$, $\|z\|_\infty$ is within a factor $1 + \epsilon$ of $\|y\|_\infty$ with probability at least $1 - o(1)$. Hence, using the arguments from [23], the median of r (independent) values gives a $1 + O(\epsilon)$ approximation to $M/\ln 2$ (the factor $\ln 2$ appears because the median of an exponentially-distributed variable is $\ln 2$).

The total dimension becomes $O(n^{1-2/k} \log n \cdot \epsilon^{-4})$.

Randomness. As described, the sketch from Theorem 1.2 requires a lot of randomness, namely $O(n)$ random variables. Reducing this number is especially important in the streaming applications, where the sketching function f itself has to be stored explicitly.

There are two main uses of independent random variables: 1) n exponentially distributed random variables u_i, and 2) the hash function $h : [n] \to [m]$ together with the variables σ_i. The second component is a common issue and can be dealt with with standard techniques: in particular, it is enough to take $q = O(\log n)$ wise independent random variables (see such details in, e.g., [24], [25], [13, Section 6], [26, Section 7]). The more tricky issue is the use of random variables u_i, for which we conjecture $O(\log n)$-independence also suffices. Independently of this conjecture, it is nonetheless possible to reduce the description complexity of f to be comparable to the size of the sketch using a different technique, namely the pseudo-random number generator of Nisan and Zuckerman [27]. In particular, assuming that the each dimension of $f(x)$ is stored using B bits, the total size of sketch is mB bits. Since m is polynomially related to n, the generator of [27] will use only $O(mB)$ truly random bits, from which it generates the variables u_i's to be used by the function f. The guarantees of the pseudo-random number generator imply that substituting the truly independent u_i's with the ones from the Nisan–Zuckerman generator will incur a failure probability which is only a small constant.

4. PRECISION SAMPLING LEMMA VIA MAX-STABILITY

The precursor to this work, [13], relied crucially on the Precision Sampling Lemma (PSL), which is a generic technique to estimate a sum $\sum a_i$ from weak estimates of each a_i. While the algorithm from above bypassed the PSL (for simplicity of exposition), it is also possible to simplify PSL itself by using the exponentially-distributed random variables. We present such a statement below.

Lemma 4.1 (Precision Sampling Lemma)**.** *Fix $a_i \in [0,1]$ for $i \in [n]$. Let u_i be chosen from the exponential distribution. Now, for $\epsilon > 0$, fix arbitrarily $\hat{a}_i$ satisfying the condition $|\hat{a}_i - a_i| \le \epsilon u_i$. Let the estimator be $\hat{A} = \max_i \hat{a}_i/u_i$. Then, we have that:*

- *there exists a (coupled) random variable $A = \sum_i a_i/u$ where u is also distributed exponentially, such that $|A - \hat{A}| \le \epsilon$ always.*
- *For any $p > 1$, we have that $\mathbb{E}\left[1/u_i^{1/p}\right] = O(1)$.*

Proof. Define $A = \max_{i\in[n]} a_i/u_i$. First note that $|A - \hat{A}| \le \epsilon$. Indeed, for each i, we have that $|a_i/u_i - \hat{a}_i/u_i| \le \epsilon$, and hence their max satisfies the same. Furthermore, using the stability property of the exponential distribution, we have that $A = \max_{i\in[n]} a_i/u_i$ is distributed as $\sum a_i/u$, where u is another exponentially distributed variable.

Second bullet is a straight-forward calculation (see also [13]):

$$\int_{u=0}^{\infty} 1/u^p \cdot e^{-u} \mathrm{d}u \le \int_{u=0}^{1} 1/u^p \mathrm{d}u + 1 \le O(1/(p-1)).$$

□

5. DISCUSSION

The use of "stability" of exponential distribution is similar to Indyk's use of p-stable distributions for sketching/streaming ℓ_p norms for $p \in (0, 2]$ [23]. Note that p-stable distributions do not exist for $p > 2$; so here the notion of "stability" is slightly different. In the former, one uses the fact that for random variables $v_1, \ldots v_n$, which are p-stable, we have that $\sum \lambda_i v_i$ is also p-stable with a well-controlled median (of the absolute value). In the latter case, we use the property that the max of several "stable" distributions is another one: $\max \lambda_i/u_i$ is distributed as λ/u (i.e., $1/u$ is "max-stable"). Note that this is useful for embedding into ℓ_∞. As it turns out, such a transformation does not increase the ℓ_2 norm of the vector much, allowing us to do the dimensionality reduction in ℓ_∞.

6. ACKNOWLEDGMENTS

We thank the anonymous referees for many insightful remarks, including for some of the ideas in Section 3.

7. REFERENCES

[1] William B. Johnson and Joram Lindenstrauss, "Extensions of lipshitz mapping into hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[2] Muthu Muthukrishnan, *Data Streams: Algorithms and Aplications*, Foundations and Trends in Theoretical Computer Science. Now Publishers Inc, Jan. 2005.

[3] David P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends in Theoretical Computer Science*, vol. 10, 2014.

[4] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor, "Analyzing graph structure via linear measurements," in *SODA*, 2012.

[5] Bruce M. Kapron, Valerie King, and Ben Mountjoy, "Dynamic graph connectivity in polylogarithmic worst case time," in *SODA*, 2013.

[6] Noga Alon, Yossi Matias, and Mario Szegedy, "The space complexity of approximating the frequency moments," *J. Comp. Sys. Sci.*, vol. 58, pp. 137–147, 1999, Previously appeared in STOC'96.

[7] Daniel M. Kane, Jelani Nelson, and David P. Woodruff, "On the exact space complexity of sketching small norms," in *SODA*, 2010.

[8] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun, "Near-optimal lower bounds on the multi-party communication complexity of set disjointness," in *CCC*, 2003, pp. 107–117.

[9] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *J. Comput. Syst. Sci.*, vol. 68, no. 4, pp. 702–732, 2004, Previously in FOCS'02.

[10] Piotr Indyk and David Woodruff, "Optimal approximations of the frequency moments of data streams," *STOC*, 2005.

[11] Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha, "Simpler algorithm for estimating frequency moments of data streams," in *SODA*, 2006, pp. 708–713.

[12] Morteza Monemizadeh and David Woodruff, "1-pass relative-error l_p-sampling with applications," in *SODA*, 2010.

[13] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak, "Streaming algorithms from precision sampling," in *FOCS*, 2011, Full version at `http://arxiv.org/abs/1011.1263`.

[14] Vladimir Braverman and Rafail Ostrovsky, "Recursive sketching for frequency moments," *CoRR*, vol. abs/1011.2571, 2010.

[15] Alexandr Andoni, Huy L. Nguyên, Yury Polyanskiy, and Yihong Wu, "Tight lower bound for linear sketches of moments," in *ICALP*, 2013, pp. 25–32, Full version at `http://arxiv.org/abs/1306.6295`.

[16] Sumit Ganguly, "Taylor polynomial estimator for estimating frequency moments," in *ICALP*, 2015, Full version at arXiv:1506.01442.

[17] Yi Li and David P Woodruff, "A tight lower bound for high frequency moment estimation with small error," in *RANDOM*, 2013, pp. 623–638.

[18] Omri Weinstein and David P. Woodruff, "The simultaneous communication of disjointness with applications to data streams," in *ICALP*, 2015.

[19] Vladimir Braverman, Jonathan Katzman, Charles Seidell, and Gregory Vorsanger, "An optimal algorithm for large frequency moments using $O(n^{1-2/k})$ bits," in *RANDOM*, 2014.

[20] Alexandr Andoni, "High frequency moments via max-stability," Blog post `http://windowsontheory.org/2012/10/04/high-frequency-moments-via-max-stability/`, 2012.

[21] David P Woodruff and Qin Zhang, "Subspace embeddings and \ell_p-regression using exponential random variables," in *COLT*, 2013, pp. 546–567.

[22] Hossein Jowhari, Mert Saglam, and Gábor Tardos, "Tight bounds for L_p samplers, finding duplicates in streams, and related problems," in *PODS*, 2011, pp. 49–58, Also in `http://arxiv.org/abs/1012.4889`.

[23] Piotr Indyk, "Stable distributions, pseudorandom generators, embeddings and data stream computation," *J. ACM*, vol. 53, no. 3, pp. 307–323, 2006, Previously appeared in FOCS'00.

[24] Jeanette P Schmidt, Alan Siegel, and Aravind Srinivasan, "Chernoff-hoeffding bounds for applications with limited independence," *SIAM Journal on Discrete Mathematics*, vol. 8, no. 2, pp. 223–250, 1995.

[25] Jelani Nelson, "Madalgo summer school 2015 johnson-lindenstrauss lecture notes," Available at `http://people.seas.harvard.edu/~minilek/madalgo2015/index.html`, 2015.

[26] Daniel M Kane and Jelani Nelson, "A derandomized sparse johnson-lindenstrauss transform," *arXiv preprint arXiv:1006.3585*, 2010.

[27] Noam Nisan and David Zuckerman, "Randomness is linear in space," *J. Comput. Syst. Sci.*, vol. 1, no. 52, pp. 43–52, 1996.