Tight Lower Bound for Linear Sketches of Moments

Alexandr Andoni[1], Huy L. Nguyễn[2], Yury Polyanskiy[3], and Yihong Wu[4]

[1] Microsoft Research SVC, andoni@microsoft.com
[2] Princeton U, hlnguyen@princeton.edu
[3] MIT, yp@mit.edu
[4] UIUC, yihongwu@illinois.edu

Abstract. The problem of estimating frequency moments of a data stream has attracted a lot of attention since the onset of streaming algorithms [AMS99]. While the space complexity for approximately computing the p^{th} moment, for $p \in (0, 2]$ has been settled [KNW10], for $p > 2$ the exact complexity remains open. For $p > 2$ the current best algorithm uses $O(n^{1-2/p} \log n)$ words of space [AKO11,BO10], whereas the lower bound is of $\Omega(n^{1-2/p})$ [BJKS04].
In this paper, we show a tight lower bound of $\Omega(n^{1-2/p} \log n)$ words for the class of algorithms based on linear sketches, which store only a sketch Ax of input vector x and some (possibly randomized) matrix A. We note that all known algorithms for this problem are linear sketches.

1 Introduction

One of the classical problems in the streaming literature is that of computing the p-frequency moments (or p-norm) [AMS99]. In particular, the question is to compute the norm $\|x\|_p$ of a vector $x \in \mathbb{R}^n$, up to $1 + \epsilon$ approximation, in the streaming model using low space. Here, we assume the most general model of streaming, where one sees updates to x of the form (i, δ_i) which means to add a quantity $\delta_i \in \mathbb{R}$ to the coordinate i of x.[5] In this setting, linear estimators, which store Ax for a matrix A, are particularly useful as such an update can be easily processed due to the equality $A(x + \delta_i e_i) = Ax + A(\delta_i e_i)$.

The frequency moments problem is among the problems that received the most attention in the streaming literature. For example, the space complexity for $p \le 2$ has been fully understood. Specifically, for $p = 2$, the foundational paper of [AMS99] showed that $O_\epsilon(1)$ words (linear measurements) suffice to approximate the Euclidean norm[6]. Later work showed how to achieve the same space for all

[5] For simplicity of presentation, we assume that $\delta_i \in \{-n^{O(1)}, \ldots, n^{O(1)}\}$, although more refined bounds can be stated otherwise. Note that in this case, a "word" (or measurement in the case of linear sketch — see definition below) is usually $O(\log n)$ bits.

[6] The exact bound is $O(1/\epsilon^2)$ words; since in this paper we concentrate on the case of $\epsilon = \Omega(1)$ only, we drop dependence on ϵ.

$p \in (0, 2)$ norms [Ind06,Li08,KNW10]. This upper bound has a matching lower bound [AMS99,IW03,Bar02,Woo04]. Further research focused on other aspects, such as algorithms with improved update time (time to process an update (i, δ_i)) [NW10,KNW10,Li08,GC07,KNPW11].

In constrast, when $p > 2$, the exact space complexity still remains open. After a line of research on both upper bounds [AMS99,IW05,BGKS06,MW10], [AKO11,BO10,Gan11] and lower bounds [AMS99,CKS03,BJKS04,JST11,PW12], we presently know that the best space upper bound is of $O(n^{1-2/p} \log n)$ words, and the lower bound is $\Omega(n^{1-2/p})$ bits (or linear measurements). (Very recently also, in a *restricted* streaming model — when $\delta_i = 1$ — [BO12] achieves an improved upper bound of nearly $O(n^{1-2/p})$ words.) In fact, since for $p = \infty$ the right bound is $O(n)$ (without the log factor), it may be tempting to assume that there the right upper bound should be $O(n^{1-2/p})$ in the general case as well.

In this work, we prove a tight lower bound of $\Omega(n^{1-2/p} \log n)$ for the case of *linear estimator*. A linear estimator uses a distribution over $m \times n$ matrices A such that with high probability over the choice of A, it is possible to calculate the p^{th} moment $\|x\|_p$ from the sketch Ax. The parameter m, the number of words used by the algorithm, is also called the number of measurements of the algorithm. Our new lower bound is of $\Omega(n^{1-2/p} \log n)$ measurements/words, which matches the upper bound from [AKO11,BO10]. We stress that essentially all known algorithms in the general streaming model are in fact linear estimators.

Theorem 1. *Fix $p \in (2, \infty)$. Any linear sketching algorithm for approximating the p^{th} moment of a vector $x \in \mathbb{R}^n$ up to a multiplicative factor 2 with probability 99/100 requires $\Omega(n^{1-2/p} \log n)$ measurements.*

In other words, for any $p \in (2, \infty)$ there is a constant C_p such that for any distribution on $m \times n$ matrices A with $m < C_p n^{1-2/p} \log n$ and any function $f : \mathbb{R}^{m \times n} \times \mathbb{R}^m \to \mathbb{R}_+$ we have

$$\inf_{x \in \mathbb{R}^n} \Pr\left(\frac{1}{2}\|x\|_p \le f(A, Ax) \le 2\|x\|_p\right) \le \frac{99}{100}. \tag{1}$$

The proof uses similar hard distributions as in some of the previous work, namely all coordinates of an input vector x have random small values except for possibly one location. To succeed on these distributions, the algorithm has to distinguish between a mixture of Gaussian distributions and a pure Gaussian distribution. Analyzing the optimal probability of success directly seems too difficult. Instead, we use the χ^2-divergence to bound the success probability, which turns out to be much more amenable to analysis.

From a statistical perspective, the problem of linear sketches of moments can be recast as a minimax statistical estimation problem where one observes the pair (Ax, A) and produces an estimate of $\|x\|_p$. More specifically, this is a functional estimation problem, where the goal is to estimation some functional (in this case, the p^{th} moment) of the parameter x instead of estimating x directly. Under this decision-theoretic framework, our argument can be understood as Le Cam's two-point method for deriving minimax lower bounds [LC86]. The idea is to use a binary hypotheses testing argument where two priors (distributions of x) are constructed, such that 1) the p^{th} moment of x differs by a constant factor under the respective prior; 2) the resulting distributions of the sketches Ax are indistinguishable. Consequently there exists no moment estimator which can achieve constant relative error. This approach is also known as the method of fuzzy hypotheses [Tsy09, Section 2.7.4]. See also [BL96,IS03,Low10,CL11] for the method of using χ^2-divergence in minimax lower bound.

We remark that our proof does not give a lower bound as a function of ϵ (but [Woo13] independently reports progress on this front).

1.1 Preliminaries

We use the following definition of divergences.

Definition 1. *Let P and Q be probability measures. The* χ^2-divergence *from P to Q is*

$$\chi^2(P||Q) \triangleq \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q$$
$$= \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1$$

The total variation distance *between P and Q is*

$$V(P,Q) \triangleq \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q| \tag{2}$$

The operational meaning of the total variation distance is as follows: Denote the optimal sum of Type-I and Type-II error probabilities of the binary hypotheses testing problem $H_0 : X \sim P$ versus $H_1 : X \sim Q$ by

$$\mathcal{E}(P,Q) \triangleq \inf_A \{P(A) + Q(A^{\mathrm{c}})\}, \tag{3}$$

where the infimum is over all measurable sets A and the corresponding test is to declare H_1 if and only if $X \in A$. Then

$$\mathcal{E}(P,Q) = 1 - V(P,Q). \tag{4}$$

The total variation and the χ^2-divergence are related by the following inequality [Tsy09, Section 2.4.1]:

$$2V^2(P,Q) \leq \log(1 + \chi^2(P||Q)) \tag{5}$$

Therefore, in order to establish that two hypotheses cannot be distinguished with vanishing error probability, it suffices to show that the χ^2-divergence is bounded.

One additional fact about V and χ^2 is the data-processing property [Csi67]: If a measurable function $f : A \to B$ carries probability measure P on A to P' on B, and carries Q to Q' then

$$V(P,Q) \geq V(P',Q')\,. \tag{6}$$

2 Lower Bound Proof

In this section we prove Theorem 1 for arbitrary fixed measurement matrix A. Indeed, by Yao's minimax principle, we only need to demonstrate an input distribution and show that any deterministic algorithm succeeding on this distribution with probability 99/100 must use $\Omega(n^{1-2/p}\log n)$ measurements.

Fix $p \in (2, \infty)$. Let $A \in \mathbb{R}^{m\times n}$ be a fixed matrix which is used to produce the linear sketch, where $m < C_p n^{1-2/p}\log n$ is the number of measurements and C_p is to be specified. Next, we construct distributions D_1 and D_2 for x to fulfill the following properties:

1. $\|x\|_p \leq Cn^{1/p}$ on the entire support of D_1, and $\|x\|_p \geq 4Cn^{1/p}$ on the entire support of D_2, for some appropriately chosen constant C.
2. Let E_1 and E_2 denote the distribution of Ax when x is drawn from D_1 and D_2 respectively. Then $V(E_1, E_2) \leq 98/100$.

The above claims immediately imply the desired (1) via the relationship between statistical tests and estimators. To see this, note that any moment estimator f induces a test for distinguishing E_1 versus E_2: declare D_2 if and only if $\frac{f(A,Ax)}{2Cn^{1/p}} \geq 1$. In other words,

$$\begin{aligned}
&\Pr_{x\sim\frac{1}{2}(D_1+D_2)}\left(\frac{1}{2}\|x\|_p \leq f(A,Ax) \leq 2\|x\|_p\right) \\
&\leq \frac{1}{2}\Pr_{x\sim D_2}\left(f(A,Ax) > 2Cn^{1/p}\right) + \frac{1}{2}\Pr_{x\sim D_1}\left(f(A,Ax) \leq 2Cn^{1/p}\right) \qquad (7)\\
&\leq \frac{1}{2}(1 + V(E_1,E_2)) \leq \frac{99}{100}, \qquad (8)
\end{aligned}$$

where the last line follows from the characterization of the total variation in (2).

The idea for constructing the desired pair of distributions is to use the Gaussian distribution and its sparse perturbation. Since the moment of a Gaussian random vector takes values on the entire $\mathbb{R}_+$, we need to further truncate by taking its conditioned version. To this end, let $y \sim N(0, I_n)$ be a standard normal random vector and t a random index uniformly distributed on $\{1,\ldots,n\}$ and independently of y. Let $\{e_1,\ldots,e_n\}$ denote the standard basis of $\mathbb{R}^n$. Let $\bar{D}_1$ and $\bar{D}_2$ be input distributions defined as follows: Under the distribution $\bar{D}_1$, we let the input vector x equal to y. Under the distribution $\bar{D}_2$, we add a one-sparse perturbation by setting $x = y + C_1 n^{1/p} e_t$ with an appropriately chosen constant C_1. Now we set D_1 to be $\bar{D}_1$ conditioned on the event $E = \{z : \|z\|_p \leq Cn^{1/p}\}$, i.e., $D_1(\cdot) = \frac{\bar{D}_1(\cdot\cap E)}{\bar{D}_1(E)}$, and set D_2 to be $\bar{D}_2$ conditioned on the event $F = \{z : \|z\|_p \geq 4Cn^{1/p}\}$. By the triangle inequality,

$$\begin{aligned}
V(E_1,E_2) &\leq V(\bar{E}_1,\bar{E}_2) + V(\bar{E}_1,E_1) + V(\bar{E}_2,E_2)\\
&\leq V(\bar{E}_1,\bar{E}_2) + V(\bar{D}_1,D_1) + V(\bar{D}_2,D_2)\\
&= V(\bar{E}_1,\bar{E}_2) + \Pr_{x\sim\bar{D}_1}(\|x\|_p \geq Cn^{1/p}) + \Pr_{x\sim\bar{D}_2}(\|x\|_p \leq 4Cn^{1/p}), \qquad (9)
\end{aligned}$$

where the second inequality follows from the data-processing inequality (6) (applied to the mapping $x \mapsto Ax$). It remains to bound the three terms in (9).

First observe that for any i, $\mathbb{E}[|y_i|^p] = t_p$ where $t_p = 2^{p/2}\Gamma(\frac{p+1}{2})\pi^{-1/2}$. Thus, $\mathbb{E}[\|y\|_p^p] = nt_p$. By Markov inequality, $\|y\|_p^p \geq 100nt_p$ holds with probability at most $1/100$. Now, if we set

$$C_1 = 4 \cdot (100t_p)^{1/p} + 10, \tag{10}$$

we have $(y_t + C_1 n^{1/p})^p > 4^p \cdot 100nt_p$ with probability at least $99/100$, and hence the third term in (9) is also smaller than $\frac{1}{100}$. It remains to show that $V(\bar{E}_1, \bar{E}_2) \leq 96/100$.

Without loss of generality, we assume that the rows of A are orthonormal since we can always change the basis of A after taking the measurements. Let ϵ be a constant smaller than $1 - 2/p$. Assume that $m < \frac{\epsilon}{100C_1^2} \cdot n^{1-2/p} \log n$. Let A_i denote the i^{th} column of A. Let S be the set of indices i such that $\|A_i\|_2 \leq 10\sqrt{m/n} \leq n^{-1/p}\sqrt{\epsilon \log n}/C_1$. Let $\bar{S}$ be the complement of S. Since $\sum_{i=1}^n \|A_i\|_2^2 = m$, we have $|\bar{S}| \leq n/100$. Let s be uniformly distributed on S and $\tilde{E}_2$ the distribution of $y + C_1 n^{1/p} e_s$. By the convexity of $(P, Q) \mapsto V(P, Q)$ and the fact that $V(P, Q) \leq 1$, we have $V(\bar{E}_1, \bar{E}_2) \leq V(\bar{E}_1, \tilde{E}_2) + \frac{|\bar{S}|}{n} \leq V(\bar{E}_1, \tilde{E}_2) + 1/100$. In view of (5), it suffices to show that

$$\chi^2(\tilde{E}_2 \| \bar{E}_1) \leq c \tag{11}$$

for some sufficiently small constant c. To this end, we first prove a useful fact about the measurement matrix A.

Lemma 1. *For any matrix A with $m < \frac{\epsilon}{100C_1^2} \cdot n^{1-2/p} \log n$ orthonormal rows, denote by S the set of column indices i such that $\|A_i\|_2 \leq 10\sqrt{m/n}$. Then*

$$|S|^{-2} \sum_{i,j \in S} e^{C_1^2 n^{2/p} \langle A_i, A_j \rangle} \leq 1.03 C_1^4 (n^{-2+4/p+\epsilon} m + n^{2/p-1}\sqrt{m}) + 1$$

Proof. Because $AA^T = I_m$, we have

$$\sum_{i,j \in [n]} \langle A_i, A_j \rangle^2 = \sum_{i,j \in [n]} (A^T A)_{ij}^2 = \|A^T A\|_F^2 = \operatorname{tr}(A^T A A^T A) = \operatorname{tr}(A^T A) = \|A\|_F^2 = m.$$

We consider the following relaxation: let $x_1, \ldots, x_{|S|^2} \geq 0$ where $\sum_i x_i^2 \leq C_1^4 n^{4/p} \cdot m$ and $x_i \leq \epsilon \log n$. We now upper bound $|S|^{-2} \sum_{i=1}^{|S|^2} e^{x_i}$. We have

$$
\begin{aligned}
|S|^{-2} \sum_{i=1}^{|S|^2} e^{x_i} &= |S|^{-2} \sum_{i=1}^{|S|^2} \left(1 + x_i + \sum_{j \geq 2} \frac{x_i^j}{j!} \right) \\
&\leq 1 + |S|^{-2} \sum_{i=1}^{|S|^2} x_i + |S|^{-2} \sum_{i=1}^{|S|^2} x_i^2 \sum_{j \geq 2} \frac{(\max_{i \in [n^2]} x_i)^{j-2}}{j!} \\
&\leq 1 + |S|^{-2} \sqrt{|S|^2 \sum_i x_i^2} + |S|^{-2} (C_1^4 m n^{4/p}) \left(\frac{e^{\epsilon \log n}}{(\epsilon \log n)^2} \right) \\
&\leq 1 + 1.03 C_1^2 \sqrt{m} n^{2/p-1} + 1.03 C_1^4 n^{-2+4/p+\epsilon} m.
\end{aligned}
$$

The last inequality uses the fact that $99n/100 \leq |S| \leq n$. Applying the above upper bound to $x_{(i-1)|S|+j} = C_1^2 n^{2/p} |\langle A_i, A_j \rangle| \leq C_1^2 n^{2/p} \|A_i\| \cdot \|A_j\| \leq \epsilon \log n$, we conclude the lemma.

We also need the following lemma [IS03, p. 97] which gives a formula for the χ^2-divergence from a Gaussian location mixture to a standard Gaussian distribution:

Lemma 2. *Let P be a distribution on $\mathbb{R}^m$. Then*

$$
\chi^2(N(0, I_m) * P \,||\, N(0, I_m)) = \mathbb{E}[\exp(\langle X, X' \rangle)] - 1\,,
$$

where X and X' are independently drawn from P.

We now proceed to proving an upper bound on the χ^2-divergence between $\bar{E}_1$ and $\tilde{E}_2$.

Lemma 3.

$$
\chi^2(\tilde{E}_2 \| \bar{E}_1) \leq 1.03 C_1^4 (n^{-2+4/p+\epsilon} m + n^{2/p-1} \sqrt{m})
$$

Proof. Let $p_i = 1/|S| \ \forall i \in S$ be the probability $t = i$. Recall that s is the random index uniform on the set $S = \{i \in [n] : \|A_i\|_2 \leq 10\sqrt{m/n}\}$. Note that $Ay \sim N(0, AA^T)$. Since $AA^T = I_m$, we have $\bar{E}_1 = N(0, I_m)$. Therefore $A(y + C_1 n^{1/p}) \sim \tilde{E}_2 = \frac{1}{|S|} \sum_{i \in S} N(A_i, I_m)$, a Gaussian location mixture.

Applying Lemma 2 and then Lemma 1, we have

$$
\begin{aligned}
\chi^2(\tilde{E}_2 \| \bar{E}_1) &= \sum_{i,j \in S} p_i p_j e^{C_1^2 n^{2/p} \langle A_i, A_j \rangle} - 1 \\
&\leq 1.03 C_1^4 (n^{-2+4/p+\epsilon} m + n^{2/p-1} \sqrt{m}).
\end{aligned}
$$

Finally, to finish the lower bound proof, since $\epsilon < 1 - 2/p$ we have $n^{-2+4/p+\epsilon} m + n^{2/p-1} \sqrt{m} = o(1)$, implying (11) for all sufficiently large n and completing the proof of $V(E_1, E_2) \leq 98/100$.

3 Discussions

While Theorem 1 is stated only for constant p, the proof also gives lower bounds for p depending on n. At one extreme, the proof recovers the known lower bound for approximating the ℓ_∞-norm of $\Omega(n)$. Notice that the ratio between the $\ell_{(\ln n)/\varepsilon}$-norm and the ℓ_∞-norm of any vector is bounded by e^ε so it suffices to consider $p = (\ln n)/\varepsilon$ with a sufficiently small constant ε. Applying the Stirling approximation to the crude value of C_1 in the proof, we get $C_1 = \Theta(\sqrt{p})$. Thus, the lower bound we obtain is $\Omega(n^{1-2/p}(\log n)/C_1^2) = \Omega(n)$.

At the other extreme, when $p \to 2$, the proof also gives super constant lower bounds up to $p = 2 + \Theta(\log\log n/\log n)$. Notice that ϵ can be set to $1 - 2/p - \Theta(\log\log n/\log n)$ instead of a positive constant strictly smaller than $1 - 2/p$. For this value of p, the proof gives a polylog(n) lower bound. We leave it as an open question to obtain tight bounds for $p = 2 + o(1)$.

Acknowledgments. HN was supported by NSF CCF 0832797, and a Gordon Wu Fellowship. YP's work was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

AKO11. Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms from precision sampling. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, 2011. Full version appears on arXiv:1011.1263.

AMS99. Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comp. Sys. Sci.*, 58:137–147, 1999. Previously appeared in STOC'96.

Bar02. Ziv Bar-Yossef. The complexity of massive data set computations. *PhD thesis, UC Berkeley*, 2002.

BGKS06. Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. Simpler algorithm for estimating frequency moments of data streams. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 708–713, 2006.

BJKS04. 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.*, 68(4):702–732, 2004.

BL96. L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24:2524–2535, 1996.

BO10. Vladimir Braverman and Rafail Ostrovsky. Recursive sketching for frequency moments. *CoRR*, abs/1011.2571, 2010.

BO12. Vladimir Braverman and Rafail Ostrovsky. Approximating large frequency moments with pick-and-drop sampling. *CoRR*, abs/1212.0202, 2012.

CKS03. Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *IEEE Conference on Computational Complexity*, pages 107–117, 2003.

CL11. T. T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

Csi67. I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Gan11. Sumit Ganguly. Polynomial estimators for high frequency moments. *arXiv*, 1104.4552, 2011.

GC07. Sumit Ganguly and Graham Cormode. On estimating frequency moments of data streams. In *Proceedings of the International Workshop on Randomization and Computation (RANDOM)*, pages 479–493, 2007.

Ind06. Piotr Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. *J. ACM*, 53(3):307–323, 2006. Previously appeared in FOCS'00.

IS03. Y.I. Ingster and I.A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer, New York, NY, 2003.

IW03. Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 283–290, 2003.

IW05. Piotr Indyk and David Woodruff. Optimal approximations of the frequency moments of data streams. *Proceedings of the Symposium on Theory of Computing (STOC)*, 2005.

JST11. Hossein Jowhari, Mert Saglam, and Gábor Tardos. Tight bounds for L_p samplers, finding duplicates in streams, and related problems. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 49–58, 2011. Previously `http://arxiv.org/abs/1012.4889`.

KNPW11. Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2011. A previous version appeared as ArXiv:1007.4191, `http://arxiv.org/abs/1007.4191`.

KNW10. Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching small norms. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

LC86. Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY, 1986.

Li08. Ping Li. Estimators and tail bounds for dimension reduction in l_p ($0 < p \leq 2$) using stable random projections. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008.

Low10. M. G. Low. Chi-square lower bounds. *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, pages 22–31, 2010.

MW10. Morteza Monemizadeh and David Woodruff. 1-pass relative-error l_p-sampling with applications. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

NW10. Jelani Nelson and David Woodruff. Fast manhattan sketches in data streams. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, 2010.

PW12. Eric Price and David P. Woodruff. Applications of the Shannon-Hartley theorem to data streams and sparse recovery. In *Proceedings of the 2012 IEEE International Symposium on Information Theory*, pages 1821–1825, 2012.

Tsy09. A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Verlag, New York, NY, 2009.

Woo04. David Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2004.

Woo13. David Woodruff. Personal communication. February 2013.