# CountSketches, Feature Hashing and the Median of Three

**Kasper Green Larsen** [1] **Rasmus Pagh** [2] **Jakub Tětek** [2]

## Abstract

In this paper, we revisit the classic *CountSketch* method, which is a sparse, random projection that transforms a (high-dimensional) Euclidean vector $v$ to a vector of dimension $(2t-1)s$, where $t, s > 0$ are integer parameters. It is known that a CountSketch allows estimating coordinates of $v$ with variance bounded by $\|v\|_2^2/s$. For $t > 1$, the estimator takes the median of $2t-1$ independent estimates, and the probability that the estimate is off by more than $2\|v\|_2/\sqrt{s}$ is exponentially small in $t$. This suggests choosing $t$ to be logarithmic in a desired inverse failure probability. However, implementations of CountSketch often use a small, constant $t$. Previous work only predicts a constant factor improvement in this setting. Our main contribution is a new analysis of CountSketch, showing an improvement in variance to $O(\min\{\|v\|_1^2/s^2, \|v\|_2^2/s\})$ when $t > 1$. That is, the variance decreases proportionally to $s^{-2}$, asymptotically for large enough $s$.

## 1. Introduction

CountSketch (Charikar et al., 2004) is a classic low-memory algorithm for processing a data stream in one pass. It supports estimating the number of occurrences of different data items in the stream, and can also be used for fast inner product estimation, or as a building block for finding heavy hitters (see e.g. (Woodruff, 2016)). Since its introduction, CountSketch has proved to be a strong primitive for approximate computation on high-dimensional vectors. Applications in machine learning include feature selection (Aghazadeh et al., 2018), neural network compression (Chen et al., 2015), random feature mappings (Pham & Pagh, 2013), compressed gradient optimizers (Spring et al., 2019), and multitask learning (Weinberger et al., 2009) — see section 1.5 for more details.

### 1.1. Sketch description

CountSketch works in the turnstile streaming model, where one is to maintain a sketch of a vector $v \in \mathbb{R}^d$ under updates to the entries. Concretely, the vector $v$ is given in a streaming fashion as a sequence of updates $(i_1, \Delta_1), (i_2, \Delta_2), \ldots,$ where an update $(i, \Delta)$ has the effect of setting $v_i \leftarrow v_i + \Delta$ for some $\Delta \in \mathbb{R}$.

The sketch can be stored as a matrix $A$ with $2t-1$ rows and $s$ columns — alternatively viewed as a vector of dimension $(2t-1)s$. Updates to the sketch are defined by hash functions $h_1, \ldots, h_{2t-1}$ and $g_1, \ldots, g_{2t-1}$. To initialize an empty CountSketch, we pick a 2-wise independent hash function $h_i : [d] \to [s]$ mapping entries in $v$ to columns of $A$, and a 2-wise independent hash function $g_i : [d] \to \{-1, 1\}$ mapping entries in $v$ to a random sign, each for row $i \in [2t-1]$.[1]. To process the update $(j, \Delta)$ the update algorithm sets $A_{i,h_i(j)} \leftarrow A_{i,h_i(j)} + g_i(j)\Delta$ for $i = 1, \ldots, 2t-1$. Thus entry $k$ of the $i$th row of $A$ contains the sum of all coordinates $v_j$ such that $h_i(j) = k$, with each such coordinate $v_j$ multiplied by a random sign $g_i(j)$.

### 1.2. Frequency estimation

A frequency estimation query (a.k.a. point query) asks to return an estimate of an entry $v_j$. CountSketch supports such queries by returning the median of $\{g_i(j)A_{i,h_i(j)}\}_{i=1}^{2t-1}$. The classic analysis of CountSketch shows that for each row $i$ of $A$ and entry $v_j$, the estimate $\hat{v}_j^i = g_i(j)A_{i,h_i(j)}$ has expectation $v_j$ and variance at most $\|v\|_2^2/s$. Using Chebyshev's inequality, this implies that $\Pr[|\hat{v}_j^i - v_j| \geq 2\|v\|_2/\sqrt{s}] \leq 1/4$. This is often boosted to a high probability bound by taking the median $\hat{v}_j$ of the $2t-1$ row estimates $\hat{v}_j^1, \ldots, \hat{v}_j^{2t-1}$ and using a Chernoff bound to conclude that $\Pr[|\hat{v}_j - v_j| \geq 2\|v\|_2/\sqrt{s}] \leq \exp(-\Omega(t))$. A similar, but less common, analysis based on Markov's inequality can also be used to give a bound based on the $\ell_1$ norm of $v$. More concretely, it can be shown that $\mathbb{E}[|\hat{v}_j^i - v_j|] \leq \|v\|_1/s$. This can again be combined with the Chernoff bound to conclude

*Equal contribution [1]Department of Computer Science, Aarhus University, Denmark [2]BARC, Department of Computer Science, University of Copenhagen, Denmark. Correspondence to: Kasper Green Larsen <larsen@cs.au.dk>, Rasmus Pagh <pagh@di.ku.dk>, Jakub Tětek <j.tetek@gmail.com>.

*Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] A $k$-wise independent hash function has independent and uniform random hash values when restricted to any set of up to $k$ keys.

that $\Pr[|\hat{v}_j - v_j| \geq 4\|v\|_1/s] \leq \exp(-\Omega(t))$. This latter bound has a better dependency on the number of columns (and hence space usage) but potentially a worse dependency on $v$ as $\|v\|_1 \geq \|v\|_2$ for all $v$ ($\|v\|_1$ and $\|v\|_2$ are close when $v$ consists of a few large non-zero entries).

Both of the above bounds suggest using a value of $t$ that is logarithmic in the desired failure probability. However, practitioners rarely use more than a small constant number of rows, such as 3 or 5 ($t = 2, 3$) rows. Based on the classic analysis of CountSketch, this only changes the failure probability by a constant factor and has no asymptotic benefits. Nonetheless, we show in experiments (in Section 4) that already 3 rows seems to have a profound impact on the variance of the estimates. The result of one experiment is seen in Figure 1. Here the ratio between the variance with 1 and 3 rows is more than 200 when using $s = 512$ columns.

We explain these observations through new theoretical insights about CountSketch. Concretely, we prove:

**Theorem 1.** *CountSketch with $t = 2$ (3 rows) satisfies $\mathbb{E}[(\hat{v}_j - v_j)^2] \leq \min\{3\|v\|_1^2/s^2, \frac{9}{8}\|v\|_2^2/s\}$.*

The constant $\frac{9}{8}$ can in fact be replaced by 1 (Private communication, Ahle & Beretta) but we give a significantly simpler proof with the slightly worse constant.

The second term in the $\min$ is a standard bound when not using median ($t = 1$). We prove that using the median trick does not worsen this bound significantly. However, the main contribution of Theorem 1 is the bound in terms of $\|v\|_1$. Quite interestingly, the bound in terms of $\|v\|_1$ is not true if using just a single row. To see this, consider any vector $v$ with a single non-zero entry $v_i$. The estimate for any other entry $v_j$ then equals 0 with probability $1 - 1/s$ ($h(i) \neq h(j)$) and it equals $v_i g(i) g(j)$ with probability $1/s$. One therefore has $\mathbb{E}[(\hat{v}_j - v_j)^2] = v_i^2/s = \|v\|_1^2/s$. This shows that using just three rows instead of a single row effectively reduces the variance of CountSketch by a factor $s$ in terms of $\|v\|_1$. We find this new insight into one of the most fundamental sketching techniques surprising. We also show that taking the median of three asymptotically reduces the fourth moment of the error in terms of $\|v\|_2$:

**Theorem 2.** *CountSketch with $t = 2$ (3 rows) satisfies $\mathbb{E}[(\hat{v}_j - v_j)^4] \leq 3\|v\|_2^4/s^2$.*

If we consider the same example as above with a vector $v$ with just a single non-zero entry $v_i$, we again see that when estimating any $v_j$ with $j \neq i$ we have $\mathbb{E}[(\hat{v}_j - v_j)^4] = v_i^4/s = \|v\|_2^4/s$. Thus using $t = 2$ (3 rows) rather than $t = 1$ (1 row) reduces the fourth moment by a factor $s$ in terms of $\|v\|_2$. We find it quite remarkable that a constant factor increase in the number of rows increases the utilization of the number of columns by a linear factor both in terms of the variance as a function of $\|v\|_1$ and the fourth moment as a function of $\|v\|_2$. Combined with our experiments, this strongly suggest that one should always use at least 3 rows in practice. We extend our results to any $t$ and show:

**Theorem 3.** *CountSketch with median of $2t-1$ rows satisfies $\mathbb{E}[|\hat{v}_j - v_j|^t] \leq 2^{2t-1}\|v\|_1^t/s^t$ and $\mathbb{E}[(\hat{v}_j - v_j)^{2t}] \leq 2^{2t-1}\|v\|_2^{2t}/s^t$.*

Thus we can bound the $t$th moment optimally (up to the $2^{2t-1}$ factor) in terms of $\|v\|_1$ and similarly for the $2t$'th moment in terms of $\|v\|_2$.

### 1.3. Inner product estimation

Another use case of CountSketch is in fast inner product estimation. Concretely, given two vectors $v, w \in \mathbb{R}^d$, if one builds a CountSketch on both vectors using *the same* random hash functions $h_1, \ldots, h_{2t-1}$ and $g_1, \ldots, g_{2t-1}$ (i.e. the same seeds), then one can quickly estimate $\langle v, w \rangle$ from the two sketches. More precisely, let $A^v$ and $A^w$ denote the matrices constructed for $v$ and $w$, respectively. For any row $i$, the inner product $\langle A_i^v, A_i^w \rangle = \sum_{j=1}^s A_{i,j}^v A_{i,j}^w$ is an unbiased estimator of $\langle v, w \rangle$. Moreover, one can show that $\mathbb{E}[(\langle A_i^v, A_i^w \rangle - \langle v, w \rangle)^2] \leq 2\|v\|_2^2\|w\|_2^2/s$ if we replace $g$ by a 4-wise independent hash function (rather than just 2-wise). Combining this with Chebyshev's inequality yields

$$\Pr[|\langle A_i^v, A_i^w \rangle - \langle v, w \rangle| > (2\sqrt{2})\|v\|_2\|w\|_2/\sqrt{s}] < 1/4.$$

Finally, as with frequency estimation (point queries), one can take the median over the $2t-1$ row estimates and apply a Chernoff bound to guarantee that the final estimate, denote it $X$, satisfies

$$\Pr[|X - \langle v, w \rangle| > (2\sqrt{2})\|v\|_2\|w\|_2/\sqrt{s}] < \exp(-\Omega(t)).$$

CountSketch with just a single row, $t = 1$, is in fact identical to the popular *feature hashing* scheme (Weinberger et al., 2009). Previous work has not shown any asymptotic benefits of taking the median of a small constant number of rows, using e.g. $t = 2$ or $t = 3$. Our contribution is new bounds on the variance of such inner product estimates:

**Theorem 4.** *For two vectors $v, w \in \mathbb{R}^d$, let $A^v$ and $A^w$ denote the two matrices representing a CountSketch of the two vectors when using the same random hash functions, where the $g_i$ are 4-wise independent. Let $X$ denote the median of $\langle A_i^v, A_i^w \rangle$ over rows $i = 1, \ldots, 2t-1$. Then CountSketch with $t = 2$ satisfies*

$$\mathbb{E}[(X - \langle v, w \rangle)^2] \leq \min\{3\|v\|_1^2\|w\|_1^2/s^2, \frac{9}{4}\|v\|_2^2\|w\|_2^2/s\},$$

*and for $t > 2$:*

$$\mathbb{E}[|X - \langle v, w \rangle|^t] \leq 2^{2t-1}\|v\|_1^t\|w\|_1^t/s^t, \text{ and}$$

$$\mathbb{E}[(X - \langle v, w \rangle)^{2t}] \leq 4^{2t-1}\|v\|_2^{2t}\|w\|_2^{2t}/s^t \ .$$

*Similarly to CountSketch, the constant $\frac{9}{4}$ can be improved to 2.*

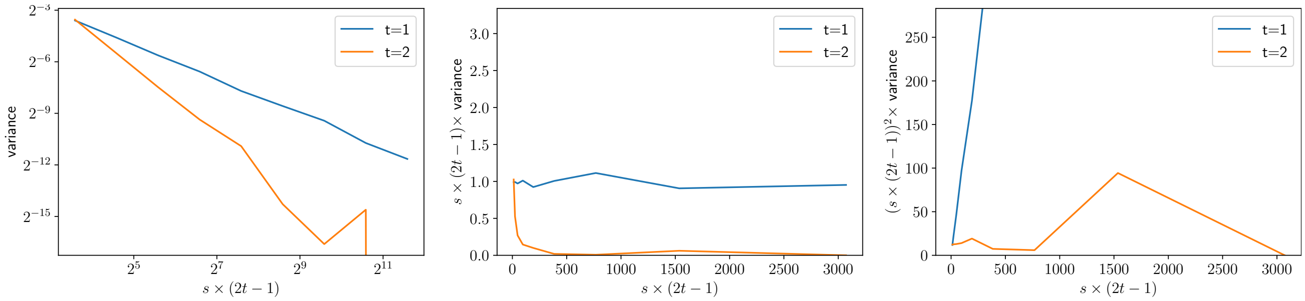

Figure 1: Variance plot of frequency estimation (point queries) for CountSketch with $t = 1$ and $t = 2$, run on a one-hot vector $v$ with a single nonzero coordinate $v_i = 1$. The x-axis shows the total space usage of $(2t-1)s$ as $s$ is increased. The first figure shows that the variances behave linearly on a log-log plot, suggesting that the variances decrease polynomially with the number of columns $s$. The second plot shows variance multiplied by $(2t-1)s$. CountSketch with $t = 1$ becomes near-constant, suggesting that its variance grows as $1/s$ (as $t$ is fixed). The third plot shows variance multiplied by $((2t-1)s)^2$ and suggests that the variance for $t = 2$ grows roughly as $1/s^2$.

We note that the bounds in terms of $\|v\|_1^2$ and $\|w\|_1^2$ can be shown only assuming 2-wise independence of the $g_i$. As with frequency estimation queries, a simple example demonstrates that the variance bound in terms of $\|v\|_1^2\|w\|_1^2$ is false for $t = 1$. Concretely, let $v$ have a single coordinate $v_i$ that is non-zero and let $w$ have a single coordinate $w_j$ with $j \neq i$ that is non-zero. Then $\langle v, w\rangle = 0$, yet the probability that $v_i$ and $w_j$ hash to the same entry is $1/s$. In that case, the estimate is either $v_i w_j = \|v\|_1\|w\|_1$ or $-v_i w_j$. This implies that $\mathbb{E}[(X - \langle v, w\rangle)^2] = \|v\|_1^2\|w\|_1^2/s$, i.e. a factor $s$ worse than the guarantees with three rows.

We have also performed experiments estimating the variance on real-world data sets, see Section 4. When $s$ is large enough (so that $\|v\|_1^2\|w\|_1^2/s^2$ becomes the smallest term), these experiments support our theoretical findings as with the frequency estimation queries.

**Discussion.** Similarly to the frequency estimation queries, our new theoretical bounds and supporting experiments strongly advocates taking the median of at least 3 rows when using CountSketch for inner product estimation. Equivalently, when using feature hashing for inner product estimation, one should take the median of at least 3 independent instantiations. This reduces the variance by a linear factor in the number of columns/coordinates of the sketch. We remark that taking the median might not be allowed in all applications. For instance, when using CountSketch/feature hashing as preprocessing for Support Vector Machines, using one row corresponds to a kernel function, while this is not the case when taking the median of multiple row estimates. The median of three can thus not be directly used in this setting.

### 1.4. New bounds on moments of the median

We prove our new variance and moment bounds for CountSketch by showing general theorems relating moments of the median of i.i.d. random variables to smaller moments of the individual random variables. These new bounds are very natural and should have applications besides in CountSketch. Moreover, we show that they are asymptotically optimal.

**Theorem 5.** *Let $X_1, \cdots, X_{2t-1}$ be $2t-1$ i.i.d. real-valued random variables and let $Y$ denote their median. For all positive integers $q$ it holds that*

$$\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}] \leq \binom{2t-1}{t} \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t \ .$$

*In particular,* $\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}] \leq 2^{2t-1} \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t$.

In many data science applications, the $X_i$ would be unbiased estimators of some desirable function of a data set, such as e.g. the coordinate $v_i$ in a vector $v$. Theorem 5 thus gives a bound on the $tq$'th moment of the estimation error of the median $Y$ in terms of just the $q$'th moment of a single variable. We remark that the median of $2t-1$ unbiased estimators is not necessarily itself an unbiased estimator, thus the bound on $\mathbb{E}[(Y - \mathbb{E}[X_1])^{tq}]$ is much more desirable than a bound on e.g. $\mathbb{E}[(Y - \mathbb{E}[Y])^{tq}]$ as the mean of $Y$ might be tricky to prove an exact bound for. However, one can, in fact, derive a bound on the variance of $Y$ itself (on $\mathbb{E}[(Y - \mathbb{E}[Y])^2)$ directly from Theorem 5:

**Corollary 1.** *Let $X_1, X_2, X_3$ be i.i.d. real-valued random variables and let $Y$ denote their median. Then*

$$\mathrm{Var}(Y) \leq \mathbb{E}[(Y - \mathbb{E}[X_1])^2] \leq 3 \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|]^2 \ .$$

*Proof.* From Theorem 5 with $q = 1$ we have

$$\mathbb{E}[(Y - \mathbb{E}[X_1])^2] \leq 3 \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|]^2.$$

Moreover, the minimizing value $\mu$ for the function $\mu \mapsto \mathbb{E}[(Y - \mu)^2]$ is the mean $\mu = \mathbb{E}[Y]$. Therefore we have $\mathrm{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \leq \mathbb{E}[(Y - \mathbb{E}[X_1])^2] \leq 3 \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|]^2$. □

In this paper, we mainly consider the case $t = 2$ with 3 rows — or equivalently 3 i.i.d. random variables.

We also prove a new inequality which we use to give bounds for the error's second moment of CountSketch in terms of $\|v\|_2$. Specifically, we show that having three i.i.d. random variables $X_1, X_2, X_3$, taking median increases the second moment of deviation from $E[X_1]$ at most by a factor of $\frac{9}{8}$. In fact, it can be proven that the second moment does not decrease at all (Private communication, Ahle & Beretta); we give a considerably simpler proof showing the factor $9/8$. Specifically, we show that

**Theorem 6.** *Let $X_1, X_2, X_3$ be i.i.d. random variables. Let $Y$ be the median of these three random variables. Then $E[(Y - E[X_1])^2] \leq \frac{9}{8} Var(X_1)$.*

### 1.5. Related work

CountSketch was originally proposed in (Charikar et al., 2004) as a method for finding heavy hitters (i.e., frequently occurring elements) in a data stream. Though there are better methods for finding heavy hitters in insertion-only data streams, CountSketch has the advantage that it is a *linear sketch*, meaning that sketches can be subtracted to form a sketch of the difference of the two vectors. It is known to be space-optimal for the problem of finding approximate $L_p$ heavy hitters in the turnstile streaming model, where both positive and negative frequency updates are possible (Jowhari et al., 2011).

**Analysis by Minton and Price.** An improved analysis of the error distribution of CountSketch was given in (Minton & Price, 2014), building on the work of (Jowhari et al., 2011). The analysis gives non-trivial bounds only when $t$ is a sufficiently large (unspecified) constant, and the exposition focuses on the case $t = \Theta(\log n)$, where $n$ is the dimension of the vector $v$. Their stated error bounds are incomparable to ours, since they are expressed in terms of (residual) $L_2$ norm of $v$.

The reader may wonder if it is possible to derive our results from the analysis in (Minton & Price, 2014). Their error bound for CountSketch is based on $\|v_{\overline{[k]}}\|_2$, where $\|v_{\overline{[k]}}\|_2$ is $v$ with the largest $k$ entries set to 0. More concretely, it is shown that for a single row of CountSketch, it holds that $\Pr[(\hat{v}_j^i - v_j)^2 > c_0 \|v_{\overline{[c_1 s]}}\|_2^2/s] < 1/4$ for some constants $c_0, c_1$. The crucial observation is that all entries of $v_{\overline{[c_1 s]}}$ are bounded by $\|v\|_1/(c_1 s)$ and therefore one has $\|v_{\overline{[c_1 s]}}\|_2^2 = O(\|v\|_1 \|v\|_1/s)$. Inserting this gives $\Pr[(\hat{v}_j^i - v_j)^2 > c_2 \|v\|_1^2/s^2] < 1/4$ and this may be combined with Chernoff bounds to give high probability bounds for the median of multiple rows in terms of $\|v\|_1$. Already with one row, this looks similar to our bound on the variance of the median of 3 rows (Theorem 1) which stated that $\mathbb{E}[(\hat{v}_j - v_j)^2] \leq 3\|v\|_1^2/s^2$. However, as our counterexample above suggests, there is no way of extending the ideas of (Minton & Price, 2014) to prove $\mathbb{E}[(\hat{v}_j^i - v_j)^2] = O(\|v\|_1^2/s^2)$ as it is simply false for $t = 1$. Indeed the way (Minton & Price, 2014) proves their bound is by analysing the $c_1 s$ largest entries separately from the remaining entries, bounding $\mathbb{E}[(\hat{v}_j^i - v_j)^2]$ only for the small entries in $v_{\overline{[c_1 s]}}$. Thus our new variance bounds do not follow from their work.

The experiments in (Minton & Price, 2014) focus on the setting where $t$ is relatively large, with 20 or 50 rows, i.e., about an order of magnitude larger space usage than we have for $t = 2$.

**Dimension reduction.** CountSketch can be used as a *dimensionality reduction* technique that is simpler and more computationally efficient than the classical Johnson-Lindenstrauss embedding (Johnson & Lindenstrauss, 1984). In this setting there is no estimator, the sketch vector is simply considered a vector in $(2t - 1)s$ dimensions. Generalized versions of CountSketch have been shown to yield a time-accuracy trade-off (Dasgupta et al., 2010; Kane & Nelson, 2014).

In machine learning, a variant of CountSketch, now known as *feature hashing*, was independently introduced in (Weinberger et al., 2009), focusing on applications in multitask learning. Feature hashing reduces variance in a slightly different way than CountSketch, by initially increasing the dimension of the input vector by a factor $t$ in a way that preserves $L_2$ distances exactly but reduces the $L_\infty$ norm of vectors by a factor $\sqrt{t}$. In (Chen et al., 2015), CountSketch/feature hashing was wired into the architecture of a neural network in order to reduce the number of model parameters (without the use of medians). CountSketch has also been used in the construction of *random feature mappings* (Pham & Pagh, 2013; Ahle et al., 2020), which can be seen as dimension-reduced versions of explicit feature maps.

**Further machine learning applications.** CountSketch, with the median estimator, has been used in several machine learning applications. In (Aghazadeh et al., 2018), CountSketch was used with $t = 2$ (3 rows) for large-scale feature selection. In (Spring et al., 2019), CountSketch was used for compressing gradient optimizers in stochastic gradient descent. The related *count-min* sketch (Cormode & Muthukrishnan, 2005), which is the special case of CountSketch where we fix $g(x) = 1$, is a popular choice in applications where vectors have non-negative entries. The count-min estimator takes advantage of non-negativity by taking the minimum of $t$ estimates, and the error distribution can be analyzed in terms of the $L_1$ norm of $v$. We note that a count-min sketch with a fully random hash function can be used to simulate a CountSketch with $s/2$ entries computing the pairwise difference of entries whose index differ in the last bit (effectively using the least significant bit as the hash function $g$).

## 2. Moments of the Median

In this section, we prove our new inequalities for moments of the median. We first state and prove an integral inequality which the proof of the theorem relies on.

**Lemma 1.** *Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-increasing function and let t be a positive integer. Then*

$$\int_0^\infty f(\sqrt[t]{x})^t dx \le \Big(\int_0^\infty f(x)dx\Big)^t \ .$$

*Proof.* Since the function is non-increasing, it is measurable. Moreover, since it is non-negative, the integrals are defined (possibly equal to $+\infty$). We have:

$$\begin{aligned}
&\Big(\int_0^\infty f(x)dx\Big)^t \\
&\quad= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^t f(x_i)dx_1 \dots dx_t \\
&\quad= t! \int_0^\infty \int_0^{x_t} \cdots \int_0^{x_2} \prod_{i=1}^t f(x_i)dx_1 \dots dx_t && (1)\\
&\quad\ge t! \int_0^\infty \int_0^{x_t} \cdots \int_0^{x_2} f(x_t)^t dx_1 \dots dx_t && (2)\\
&\quad= t! \int_0^\infty f(x_t)^t \int_0^{x_t} \cdots \int_0^{x_2} 1 dx_1 \dots dx_t \\
&\quad= t! \int_0^\infty f(x_t)^t \frac{x_t^{t-1}}{(t-1)!} dx_t && (3)\\
&\quad= \int_0^\infty f(x)^t t x^{t-1} dx = \int_0^\infty f(\sqrt[t]{x})^t dx \ .
\end{aligned}$$

The integral in (1) is exactly over the set $0 \le x_1 \le x_2 \le \cdots \le x_t$. There are $t!$ such sets, each determined by an ordering of the variables. Since $\prod_{i=1}^t$ is a symmetric function (by commutativity) it integrates to the same value over each of these sets. Moreover, these sets partition the set $[0,\infty)^t$ (up to a set of measure 0 corresponding to when two variables are equal). Since we have a partition into $t!$ sets and the integral over each set from the partition is the same, the integral over each set is a $t!$-fraction of the integral over the whole space, and (1) holds. (2) holds because $f$ is non-increasing and $x_1 \le x_2, \cdots, x_t$. (3) holds because the inner integrals correspond to the volume of the $t-1$-dimensional ordered simplex scaled by a factor of $x_t$ and the volume of $t-1$-dimensional ordered simplex is $\frac{1}{(t-1)!}$ (this holds by symmetry, and can be argued the same way as (1)). The final equality holds by substituting $x = x^t$. □

**Restatement of Theorem 5.** *Let $X_1, \cdots, X_{2t-1}$ be $2t-1$ i.i.d. real-valued random variables and let $Y$ denote their median. For all positive integers $q$ it holds that*

$$\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}] \le \tbinom{2t-1}{t} \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t \ .$$

*In particular,* $\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}] \le 2^{2t-1} \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t$.

*Proof.* Notice that since $Y$ is the median of $X_1, \dots, X_{2t-1}$ and the $X_i$'s have the same mean, we can only have $|Y - \mathbb{E}[X_1]|^{tq} \ge x$ when at least $t$ variables $X_i$ have $|X_i - \mathbb{E}[X_i]|^{tq} \ge x$. There are $\binom{2t-1}{t}$ choices for such $t$ variables, so by the union bound, independence and identical distribution of the $X_i$'s, we have for any $x$ that:

$$\Pr[|Y - \mathbb{E}[X_1]|^{tq} \ge x] \le \tbinom{2t-1}{t} \Pr[|X_1 - \mathbb{E}[X_1]|^{tq} \ge x]^t.$$

We can thus bound $\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}]$ as:

$$\begin{aligned}
&\mathbb{E}[|Y - \mathbb{E}[X_1]|^{tq}] \\
&\quad= \int_0^\infty \Pr[|Y - \mathbb{E}[X_1]|^{tq} \ge x]dx \\
&\quad\le \tbinom{2t-1}{t} \int_0^\infty \Pr[|X_1 - \mathbb{E}[X_1]|^{tq} \ge x]^t dx \\
&\quad= \tbinom{2t-1}{t} \int_0^\infty \Pr[|X_1 - \mathbb{E}[X_1]|^{q} \ge \sqrt[t]{x}]^t dx \\
&\quad\le \tbinom{2t-1}{t} \Big(\int_0^\infty \Pr[|X_1 - \mathbb{E}[X_1]|^{q} \ge x]dx\Big)^t \\
&\quad= \tbinom{2t-1}{t} \cdot \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t,
\end{aligned}$$

where the first and last equalities hold by a standard identity for non-negative random variables, and the last inequality holds by Lemma 1 since $\Pr[|X_1 - \mathbb{E}[X_1]|^q \ge x]$ is a non-increasing non-negative function. □

The bound shown in this section can easily be seen to be asymptotically optimal. Consider $X_i$'s which take value $k$ with probability $1/k$ and are zero otherwise. Then

$$\begin{aligned}
&\mathbb{E}[|Y - \mathbb{E}[X_1]|^{qt}] \\
&\ =(k-1)^{qt} \Pr[Y = k] \\
&\ =(k-1)^{qt} \Big(\binom{2t-1}{t} \Pr[X_1 = k]^t + O(\Pr[X_1 = k]^2 t)\Big) \\
&\sim \binom{2t-1}{t} \frac{(k-1)^{qt}}{k^t} \\
&\sim \binom{2t-1}{t} (k-1)^{(q-1)t}
\end{aligned}$$

where the limit in $\sim$ is taken for $k \to \infty$. The second equality holds by the inclusion-exclusion principle. On the other hand, the bound given by our theorem is

$$\binom{2t-1}{t} \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^q]^t$$

$$
\begin{aligned}
&= \binom{2t-1}{t} \left(\frac{1}{k}(k-1)^q\right)^t \\
&\sim \binom{2t-1}{t} (k-1)^{(q-1)t}
\end{aligned}
$$

We now prove Theorem 6.

**Restatement of Theorem 6.** *Let $X_1, X_2, X_3$ be i.i.d. random variables. Let $Y$ be the median of these three random variables. Then $E[(Y - E[X_1])^2] \le \frac{9}{8} Var(X_1)$.*

*Proof.* We can assume without loss of generality that $E[X_1] = 0$. Then

$$
\begin{aligned}
E[Y^2] &= \int_0^\infty \Pr[Y^2 \ge x] dx \\
&\le \int_0^\infty \Pr[|\{i, \text{ s.t. } X_i^2 \ge x\}| \ge 2]
\end{aligned}
$$

Let $p_x = \Pr[X_1^2 \ge x]$. Then $\Pr[|\{i \text{ s.t. } X_i^2 \ge x\}| \ge 2] \le 3p_x^2(1-p_x) + p_x^3$. It holds for any $z \ge 0$ that $3z^2(1-z) + z^3 \le \frac{9}{8} z$. We can, therefore, bound the integral by

$$
\begin{aligned}
&\le \int_0^\infty \frac{9}{8} \Pr[X_1^2 \ge x] \\
&= \frac{9}{8} E[X_1^2] \\
&= \frac{9}{8} Var(X_1)
\end{aligned}
$$

□

## 3. CountSketch

In this section, we prove our new bounds on the variance (Theorem 1) and 4th moment (Theorem 2) for CountSketch with 3 rows ($t = 2$) as well as our general theorem with the median of $2t - 1$ estimates (Theorem 5).

**Frequency estimation.** Recall that CountSketch with three rows computes an estimate $\hat{v}_j^i$ for each of three rows $i = 1, 2, 3$ and returns the median $\hat{v}_j$ as its estimate of $v_j$. From Theorem 5, we see that to obtain variance and 4th moment bounds for $\hat{v}_j$, we only need to bound $\mathbb{E}[|\hat{v}_j^1 - \mathbb{E}[\hat{v}_j^1]|^q]$ for $q = 1, 2$. Such a bound follows from existing work, see e.g. (Cormode & Yi, 2020), Fact 3.4:

**Lemma 2.** *CountSketch satisfies $\mathbb{E}[\hat{v}_j^1] = v_j$, $\mathbb{E}[|\hat{v}_j^1 - v_j|] \le \|v\|_1/s$ and $\mathbb{E}[(\hat{v}_j^1 - v_j)^2] \le \|v\|_2^2/s$.*

Theorem 1 follows by instantiating Theorem 5 with $q = 1$ and the facts $\mathbb{E}[\hat{v}_j^1] = v_j$, $\mathbb{E}[|\hat{v}_j^1 - v_j|] \le \|v\|_1/s$ from Lemma 2. Theorem 2 follows by instantiating Theorem 5 with $q = 2$ and the facts $\mathbb{E}[\hat{v}_j^1] = v_j$, $\mathbb{E}[(\hat{v}_j^1 - v_j)^2] \le \|v\|_2^2/s$ from Lemma 2. Finally, Theorem 3 also follows as an immediate corollary of Theorem 5 and Lemma 2.

**Inner product estimation.** Similarly to the case of frequency estimation (point queries), we prove our new guarantees in Theorem 4 by invoking our general theorems on moments of the median. All we need is moment bounds for a single row. The following is more or less standard. We include the proof in the full version (Larsen et al., 2021).

**Lemma 3.** *For two vectors $v, w \in \mathbb{R}^d$, let $A^v$ and $A^w$ denote the two matrices representing a CountSketch of the two vectors when using the same random hash functions. Then $\mathbb{E}[\langle A_1^v, A_1^w \rangle] = \langle v, w \rangle$ and $\mathbb{E}[|\langle A_1^v, A_1^w \rangle - \langle v, w \rangle|] \le \|v\|_1 \|w\|_1/s$. Moreover, if $g$ is 4-wise independent, then we also have $\mathbb{E}[(\langle A_1^v, A_1^w \rangle - \langle v, w \rangle)^2] \le 2\|v\|_2^2 \|w\|_2^2/s$.*

Theorem 4 follows by combining Lemma 3 and Theorem 5.

## 4. Experiments

In this section, we empirically support our new theoretical bounds by estimating the variance of CountSketch with 1 row and 3 rows on different data sets. We implemented CountSketch in C++ using the multiply-shift hash function (Dietzfelbinger, 1996) as the 2-wise independent hash functions $h$ and $g$. We seeded the hash functions using random numbers generated using the built-in Mersenne twister 64-bit pseudorandom generator. Experiments were run both for frequency estimation (Section 4) and for inner product estimation (Section 4).

**Frequency estimation.** We ran experiments on two real-world data sets and two synthetic data sets. The real-world data sets come in the form of a stream of items, with the same item occurring multiple times. Instead of running numerous $(i, 1)$ updates ($v_i \leftarrow v_i + 1$), we have simply computed the number of occurrences $c_i$ of each item. We then normalize the occurrences $c_i \leftarrow c_i / \sum_j c_j$ to obtain unit $\ell_1$-norm and then run a single update $v_i \leftarrow v_i + c_i$ for each item $i$ at the end. This produces the exact same CountSketch as when processing the updates one by one (with normalization). The data sets are described in the following:

- **Kosarak:** An anonymized click-stream dataset of a Hungarian online news portal. [2] It consists of transactions, each of which has several items. We created a vector with one entry for each item, storing the total number of occurrences of that item. The vector has 41270 entries, and when normalized to have $\ell_1$-norm 1, its $\ell_2$-norm is 0.112 and the largest entry is 0.075.
- **Sentiment140:** A collection of 1.6M tweets from Twitter (Go et al., 2009). We extracted all words that occur at least twice, and created a vector with one entry per

[2] Provided by Ferenc Bodon to the FIMI data set located at http://fimi.uantwerpen.be/data/.

word, containing the total number of occurrences of that word in the tweets. The vector has 147071 entries, and when normalized to have $\ell_1$-norm 1, its $\ell_2$-norm is 0.0773 and the largest entry is 0.0382.

- **Zipfian:** The Zipfian distribution with skew $\alpha$ and $n$ items is a probability distribution where the $k$th item has probability $k^{-\alpha}/\sum_{j=1}^{n} j^{\alpha}$. Such distributions have been shown to fit a large variety of real-world data. We created two data sets with $n = 1000$ items using skews $\alpha = 0.8$ and $\alpha = 1.2$, considering the vector of probabilities. For $\alpha = 0.8$, the $\ell_2$-norm is 0.097 and the largest entry is 0.065. For $\alpha = 1.2$, the $\ell_2$-norm is 0.2713 and the largest entry is 0.231. We include results for $\alpha = 0.8$ in the full version (Larsen et al., 2021).

The results of the experiments can be seen in Figures 2-4. For each experiment, we plot the variance as a function of the total space usage $(2t - 1)s$ as we increase the number of columns $s$. We run experiments with $s = 2^2, 2^3, \ldots, 2^{10}$ on each data set. For each choice of $s$, we estimate the variance by constructing 1000 CountSketches on the input with new randomness for each. For each CountSketch we pick 100 random items and compute the estimation error for each. We sum the squares of all these estimation errors and divide by $100 \times 1000$ (for small data sets with less than 5000 items, we instead build $10^6$ CountSketches and make a single estimation on each).

On all four data sets, we make three plots of the data. On the first, we show a log-log plot and observe that in all experiments, the variances look linear on the plot, supporting a polynomial dependency on $s$. Second, we scale the variances by $(2t-1)s$ and plot it on a linear scale. In all experiments, the scaled variance for $t = 1$ looks constant, supporting a $1/s$ dependency on the number of columns $s$. Third, we scale the variance by $((2t - 1)s)^2$ and plot it on a linear scale. The scaled variance for $t = 2$ looks almost constant in all experiments, supporting a $1/s^2$ dependency on the number of columns. We remark that our theoretical bound in Theorem 1 guarantees $\mathbb{E}[(\hat{v}_j - v_j)^2] \leq 3\|v\|_1^2/s^2$. Since $\|v\|_1 = 1$ in all our data sets, we expect the CountSketches with $t = 2$ on the third plots to stay below $3(2t - 1)^2 = 27$ on the y-axis, which it does in all experiments (it even stays below 4 on all but the last experiment).

Table 1 shows the variance on the different data sets using CountSketch with $s = 1024$ rows. In all cases, that increasing CountSketch parameter $t$ from 1 to 2 clearly provides major reductions in variance, ranging from a factor of about 28 to 174.

We also perform experiments measuring the 4th moment of the estimation errors. These can be found in the full version (Larsen et al., 2021).

| Data Set | Variance $t = 1$ | Variance $t = 2$ | Ratio |
|---|---|---|---|
| Kosarak | $1.25 \times 10^{-5}$ | $1.42 \times 10^{-7}$ | 88.0 |
| Sentiment140 | $5.94 \times 10^{-6}$ | $2.13 \times 10^{-7}$ | 27.9 |
| Zipfian $\alpha = 0.8$ | $9.56 \times 10^{-6}$ | $2.09 \times 10^{-7}$ | 45.7 |
| Zipfian $\alpha = 1.2$ | $6.94 \times 10^{-5}$ | $3.99 \times 10^{-7}$ | 173.9 |

Table 1: Variances for different data sets with 2 and 3 rows ($t = 1, 2$) of CountSketch. In all experiments, we consider a CountSketch with $s = 1024$ columns. The ratio in the last column of the table gives the relative difference between using 1 and 3 rows.

To summarize, we believe our empirical findings support our new theoretical bounds on the variance and 4th moment. Moreover, our results strongly suggest that practitioners use $t \geq 2$ with CountSketch as it provides major reductions in variance at little increase in time and memory efficiency.

**Inner product estimation.** In the following, we perform experiments where we use CountSketch for inner product estimation. We perform experiments on two data sets, a synthetic and a real-world data set:

- **Disjoint 64 non-zeros:** A synthetic data set with two vectors both having 64 non-zero entries each with value $1/64$. The two vectors have disjoint supports and thus inner product 0. The $\ell_2$-norm of the vectors is $1/8 = 0.125$ and the largest entry is $1/64 \approx 0.0156$.
- **News20:** A collection of newsgroup documents on different topics [3]. Each document is represented by a tf-idf vector constructed on the words occurring in the documents. We used the training part of the data set for our experiments. The data set has 11314 distinct vectors. For comparison to our theoretical bounds, we normalize the vector $v$ representing each document such that it has $\|v\|_1 = 1$. After normalization, the average $\ell_2$-norm of a document vector is 0.1235 and the average largest entry is 0.0498.

For the Disjoint 64 Non-Zeros data set, for $10^6$ iterations, we constructed a new CountSketch on the two vectors using the same random hash functions. We then computed the squared error of the estimates and averaged over all $10^6$ iterations. For the News20 data set, we run 1000 iterations where we pick new random hash functions in each iteration. In an iteration, we pick 100 random pairs of distinct vectors, build a CountSketch on both vectors in a pair, and compute the squared estimation error. We finally average over all $100 \times 1000$ pairs. Figure 5 shows the results of experiments on the Disjoint 64 Non-Zeros data set. As before, these plots fit our theoretical guarantees in Theorem 4.

Finally, we have run experiments on the News20 data

[3] http://qwone.com/~jason/20Newsgroups/

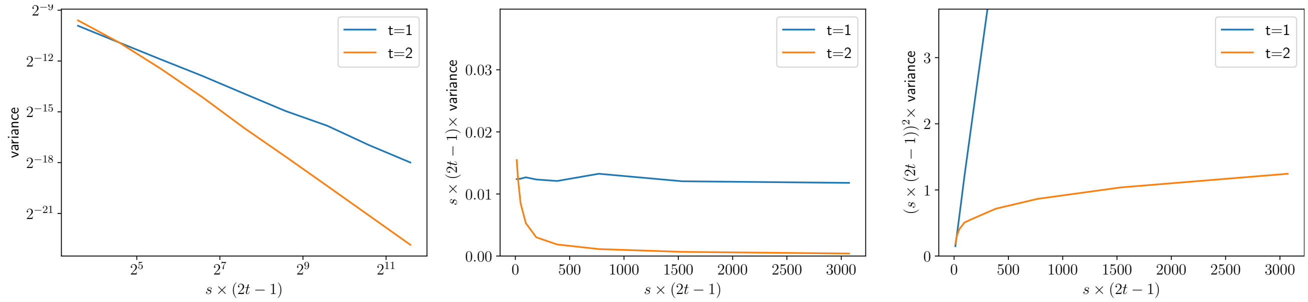

Figure 2: Variance experiments on the Kosarak data set.

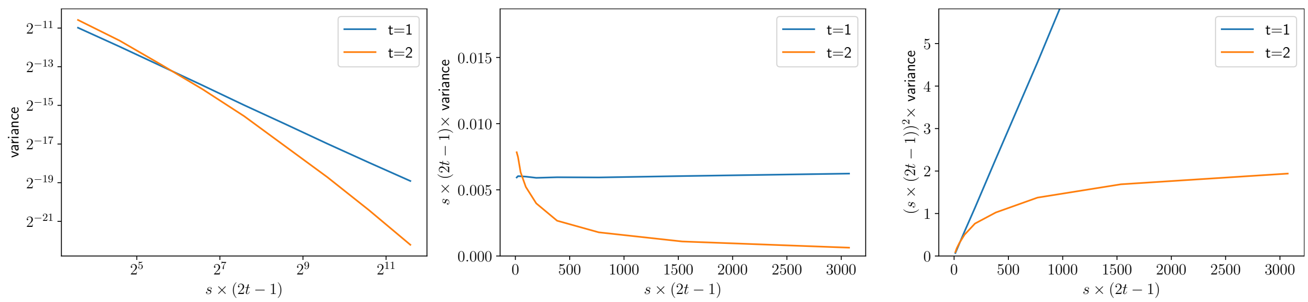

Figure 3: Variance experiments on the Sentiment140 data set.

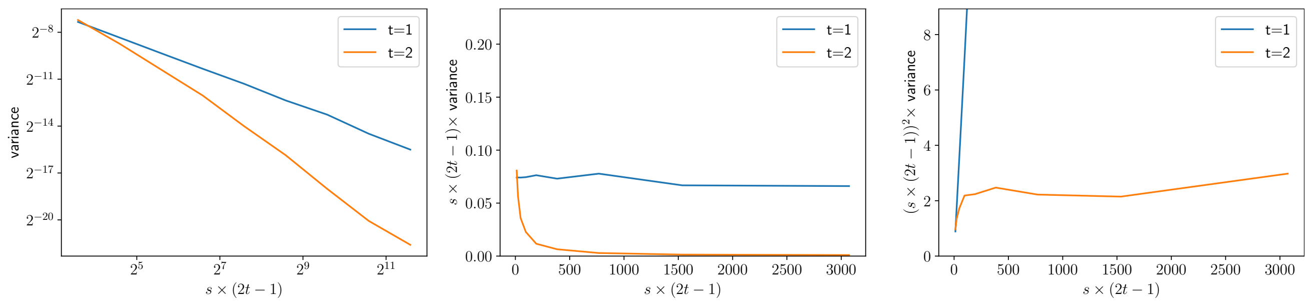

Figure 4: Variance experiments on Zipfian distribution with skew $\alpha = 1.2$.

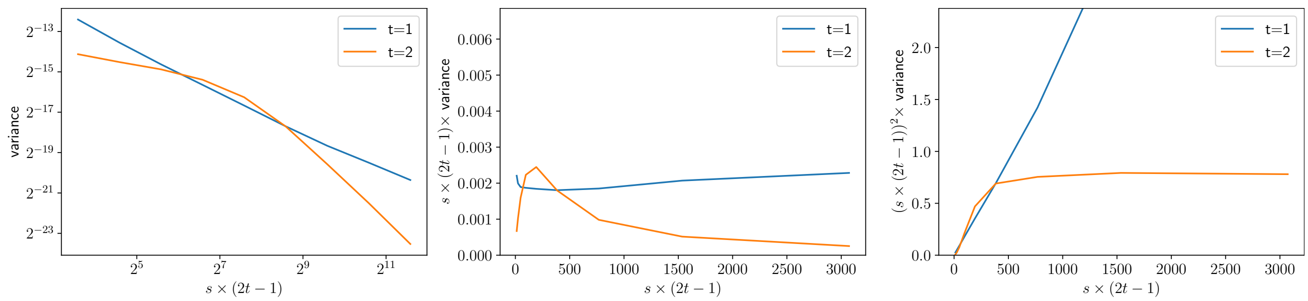

Figure 5: Variance experiments on the Disjoint 64 Non-Zeros data set.

set. The results are shown in Figure 6. Unlike in previous experiments, it appears that CountSketch with 3 rows ($t = 2$) has a variance decreasing as $1/s$, not $1/s^2$. To explain this, recall that the guarantee from Theorem 4 is $\mathbb{E}[(X - \langle v, w \rangle)^2] \leq \min\{3\|v\|_1^2 \|w\|_q^2/s^2, 2\|v\|_2^2\|w\|_2^2\}$. In the News20 data set, the average $\|v\|_2$ is 0.1235. When this is raised to the fourth power (it appears in both $\|v\|_2^2$ and $\|w\|_2^2$) it becomes very small compared to $\|v\|_1^2\|w\|_1^2 = 1$, thus the $1/s^2$ dependency should only kick in for large values of $s$. To confirm this, we have run more experiments, this time with values of $s$ ranging from $2^{10}$ to $2^{20}$. The results are shown in Figure 7.

With these larger values of $s$, we see the expected $1/s^2$ dependency in the variance for $t = 2$. To conclude on this, one may need a larger value of $s$ to see the $1/s^2$ behaviour in variance when performing inner product estimation com-

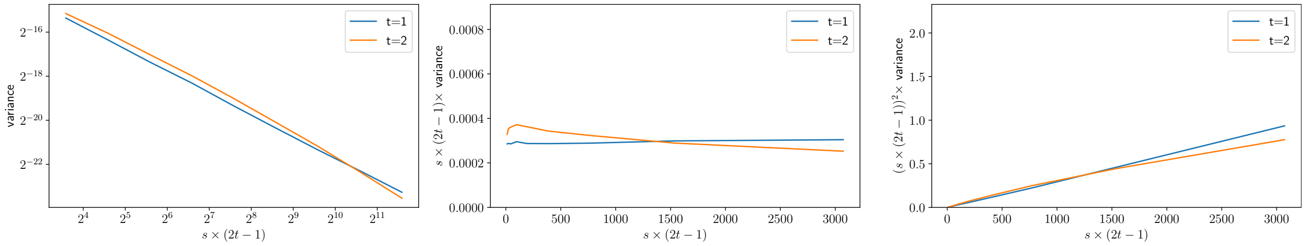

Figure 6: Variance experiments on the News20 data set.

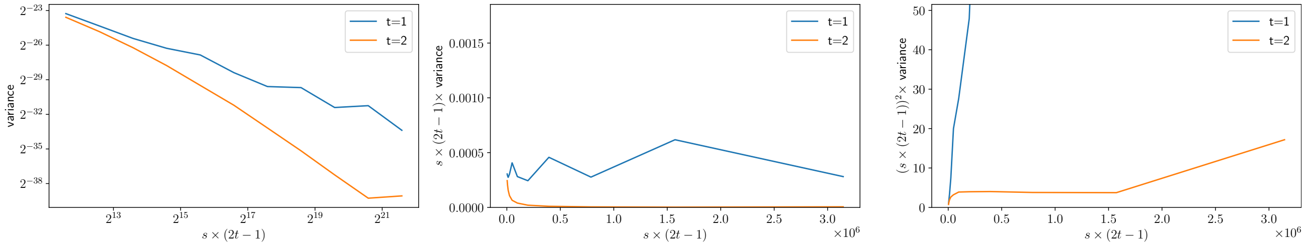

Figure 7: Variance experiments on the News20 data set and number of columns up to $s = 2^{20}$.

pared to frequency estimation. This is due to the dependency on the *product* of two vectors of either $\|v\|_1^2\|w\|_1^2$ or $\|v\|_2^2\|w\|_2^2$ compared to just the single dependency on $\|v\|_1^2$ and $\|v\|_2^2$ for frequency estimation.

As with frequency estimation, we also experimentally examine the 4th moments. These results are included in the full version (Larsen et al., 2021).

## 5. Conclusion

We have seen that taking the median of 3 estimates can significantly improve the accuracy of estimates for CountSketch. An interesting direction, that we leave open, is to take advantage of this in more applications that use CountSketch or feature hashing. A challenge is that a median operation is not available in some contexts (like neural networks, or kernel approximation), and may need to be replaced by a continuously differentiable approximation.

## Acknowledgements

We are grateful to Ora N. Fandina who has pointed out a flaw in an earlier draft. We are also grateful to Lorenzo Beretta and Thomas D. Ahle who have proved a stronger version of Theorem 6 with constant 1. Kasper Green Larsen is supported by Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 9064-00068B, a Villum Young Investigator grant and Basic Algorithms Research Copenhagen (BARC). Rasmus Pagh and Jakub Tětek are affiliated with BARC, supported by the VILLUM Foundation grant 16582. Jakub Tětek is grateful for the financial support of the Bakala Foundation.

## References

Aghazadeh, A., Spring, R., LeJeune, D., Dasarathy, G., Shrivastava, A., and Baraniuk, R. G. Mission: Ultra large-scale feature selection using count-sketches. In *Proceedings of annual International Conference on Machine Learning (ICML)*, pp. 80–88. PMLR, 2018.

Ahle, T. D. and Beretta, L. Personal communication.

Ahle, T. D., Kapralov, M., Knudsen, J. B., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of annual Symposium on Discrete Algorithms (SODA)*, pp. 141–160, 2020.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Chen, W., Wilson, J., Tyree, S., Weinberger, K., and Chen, Y. Compressing neural networks with the hashing trick.

In *Proceedings of annual International Conference on Machine Learning (ICML)*, pp. 2285–2294. PMLR, 2015.

Cormode, G. and Muthukrishnan, S. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Cormode, G. and Yi, K. *Small Summaries for Big Data*. Cambridge University Press, 2020.

Dasgupta, A., Kumar, R., and Sarlós, T. A sparse Johnson-Lindenstrauss transform. In *Proceedings of Symposium on Theory of computing (STOC)*, pp. 341–350, 2010.

Dietzfelbinger, M. Universal hashing and k-wise independent random variables via integer arithmetic without primes. In *Proceedings of Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 1046 of *Lecture Notes in Computer Science*, pp. 569–580. Springer, 1996.

Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009. URL http://help.sentiment140.com/for-students.

Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Jowhari, H., Sağlam, M., and Tardos, G. Tight bounds for lp samplers, finding duplicates in streams, and related problems. In *Proceedings of symposium on Principles of Database Systems (PODS)*, pp. 49–58, 2011.

Kane, D. M. and Nelson, J. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1), January 2014. ISSN 0004-5411. doi: 10.1145/2559902.

Larsen, K. G., Pagh, R., and Tětek, J. Countsketches, feature hashing and the median of three, 2021. URL https://arxiv.org/abs/2102.02193.

Minton, G. T. and Price, E. Improved concentration bounds for count-sketch. In *Proceedings of Annual Symposium on Discrete Algorithms (SODA)*, pp. 669–686, 2014. doi: 10.1137/1.9781611973402.51.

Pham, N. and Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of international conference on Knowledge Discovery and Data mining (KDD)*, pp. 239–247, 2013.

Spring, R., Kyrillidis, A., Mohan, V., and Shrivastava, A. Compressing gradient optimizers via count-sketches. In *Proceedings of annual International Conference on Machine Learning (ICML)*, pp. 5946–5955. PMLR, 2019.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. Feature hashing for large scale multitask learning. In *Proceedings of annual International Conference on Machine Learning (ICML)*, pp. 1113–1120, 2009.

Woodruff, D. P. New algorithms for heavy hitters in data streams (invited talk). In *International Conference on Database Theory (ICDT)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.