Exponentially Improved Dimensionality Reduction for ℓ_1: Subspace Embeddings and Independence Testing

Yi Li YILI@NTU.EDU.SG
Nanyang Technological University

David P. Woodruff DWOODRUF@CS.CMU.EDU
Carnegie Mellon University

Taisuke Yasuda TAISUKEY@CS.CMU.EDU
Carnegie Mellon University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Despite many applications, dimensionality reduction in the ℓ_1-norm is much less understood than in the Euclidean norm. We give two new oblivious dimensionality reduction techniques for the ℓ_1-norm which improve *exponentially* over prior ones:

1. We design a distribution over random matrices $\mathbf{S} \in \mathbb{R}^{r\times n}$, where $r = 2^{\text{poly}(d/(\varepsilon\delta))}$, such that given any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, with probability at least $1-\delta$, simultaneously for all $\mathbf{x}$, $\|\mathbf{SAx}\|_1 = (1\pm\varepsilon)\|\mathbf{Ax}\|_1$. Note that $\mathbf{S}$ is linear, does not depend on $\mathbf{A}$, and maps ℓ_1 into ℓ_1. Our distribution provides an exponential improvement on the previous best known map of Wang and Woodruff (SODA, 2019), which required $r = 2^{2^{\Omega(d)}}$, even for constant ε and δ. Our bound is optimal, up to a polynomial factor in the exponent, given a known $2^{\text{poly}(d)}$ lower bound for constant ε and δ.
2. We design a distribution over matrices $\mathbf{S} \in \mathbb{R}^{k\times n}$, where $k = 2^{O(q^2)}(\varepsilon^{-1}q\log d)^{O(q)}$, such that given any q-mode tensor $\mathbf{A} \in (\mathbb{R}^d)^{\otimes q}$, one can estimate the entrywise ℓ_1-norm $\|\mathbf{A}\|_1$ from $\mathbf{S}(\mathbf{A})$. Moreover, $\mathbf{S} = \mathbf{S}^1 \otimes \mathbf{S}^2 \otimes \cdots \otimes \mathbf{S}^q$ and so given vectors $\mathbf{u}_1, \ldots, \mathbf{u}_q \in \mathbb{R}^d$, one can compute $\mathbf{S}(\mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \cdots \otimes \mathbf{u}_q)$ in time $2^{O(q^2)}(\varepsilon^{-1}q\log d)^{O(q)}$, which is much faster than the d^q time required to form $\mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \cdots \otimes \mathbf{u}_q$. Our linear map gives a streaming algorithm for independence testing using space $2^{O(q^2)}(\varepsilon^{-1}q\log d)^{O(q)}$, improving the previous doubly exponential $(\varepsilon^{-1}\log d)^{q^{O(q)}}$ space bound of Braverman and Ostrovsky (STOC, 2010).

For subspace embeddings, we also study the setting when $\mathbf{A}$ is itself drawn from distributions with independent entries, and obtain a polynomial embedding dimension. For independence testing, we also give algorithms for any distance measure with a polylogarithmic-sized sketch and satisfying an approximate triangle inequality.

Keywords: Subspace embeddings, independence testing, dimension reduction

1. Introduction

Dimensionality reduction refers to mapping a set of high-dimensional vectors to a set of low-dimensional vectors while preserving their lengths and pairwise distances. A celebrated result is the Johnson-Lindenstrauss embedding, which asserts that for a random linear map $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^r$, for any fixed $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{Sx}\|_2 = (1\pm\varepsilon)\|\mathbf{x}\|_2$ with probability $1-\delta$. It is necessary and sufficient for the sketching dimension r to be $\Theta(\varepsilon^{-2}\log(1/\delta))$ Johnson and Lindenstrauss (1984); Larsen and Nelson (2017). A key property of $\mathbf{S}$ is that it is *linear* and *oblivious*, meaning that it is a linear map that does not depend on the point set. This makes it applicable in settings such as

© 2021 Y. Li, D.P. Woodruff & T. Yasuda.

the widely used streaming model, where one sees coordinates or updates to coordinates one at a time (see, e.g., Muthukrishnan (2005); Cormode et al. (2012) for surveys) and the distributed model where points are shared across servers (see, e.g., Boutsidis et al. (2016), for a discussion of different models). Here it is crucial that for points $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{S}(\mathbf{x}+\mathbf{y}) = \mathbf{S}\mathbf{x} + \mathbf{S}\mathbf{y}$, and $\mathbf{S}$ does not depend on $\mathbf{x}$ or $\mathbf{y}$. In this case, if one receives a new point $\mathbf{z}$ chosen independently of $\mathbf{S}$, then $\mathbf{S}$ still has a good probability of preserving the length of $\mathbf{z}$, whereas data-dependent linear maps $\mathbf{S}$ may change with the addition of $\mathbf{z}$, and are often slower Indyk et al. (2000). For these reasons, our focus is on linear oblivious dimensionality reduction, often referred to as "sketching".

For many problems, the 1-norm $\|\mathbf{x}\|_1 = \sum_{i=1}^n |\mathbf{x}_i|$ is more appropriate than the Euclidean norm. Indeed, this norm is used in applications demanding robustness since it is less sensitive to changes in individual coordinates. As the 1-norm is twice the variation distance between distributions, it is often the metric of choice for comparing distributions Indyk and McGregor (2008); Braverman et al. (2010); Braverman and Ostrovsky (2010a); McGregor and Vu (2015). A sample of applications involving the 1-norm includes clustering Feldman et al. (2010); Labib and Vemuri (2005), regression Clarkson (2005); Sohler and Woodruff (2011); Clarkson et al. (2013); Meng and Mahoney (2013); Woodruff and Zhang (2013); Clarkson and Woodruff (2015, 2017); Woodruff (2014), time series analysis Dodge (1992); Lawrence (2019), internet traffic monitoring Feigenbaum et al. (2002), multimodal and similarity search Aggarwal et al. (2001); Lin and Shim (1995). As stated in Aggarwal et al. (2001), "the Manhattan distance metric is consistently more preferable than the Euclidean distance metric for high dimensional data mining applications".

While useful for the Euclidean norm, the Johnson-Lindenstrauss embedding completely fails if one wants for a vector $\mathbf{x}$, that $\|\mathbf{S}\mathbf{x}\|_1 = (1 \pm \varepsilon)\|\mathbf{x}\|_1$ with probability $1 - \delta$. Indeed, the results of Wang and Woodruff Wang and Woodruff (2019) imply nearly tight bounds: for constant ε and δ, a sketching dimension of $2^{\text{poly}(n)}$ is necessary and sufficient[1]. Indyk Indyk (2006b) shows that if instead one embeds $\mathbf{x}$ into a non-normed space, namely, performs a "median of absolute values" estimator of $\mathbf{S}\mathbf{x}$, then the dimension can be reduced to $O((\log n)/\varepsilon^2)$. Such a mapping is still linear and oblivious, and this estimator is useful if one desires to approximate the norm of a single vector, for which the dimension becomes $O((\log(1/\delta))/\varepsilon^2)$ and the failure probability is δ. However, this estimator is less useful in optimization problems as it requires solving a non-convex problem after sketching. Thus, there is a huge difference in dimensionality reduction for the Euclidean and 1-norms.

This work is motivated by our poor understanding of dimensionality reduction in the 1-norm, as exemplified by two existing doubly exponential bounds for important problems: preserving a subspace of points and preserving a sum of tensor products, both of which are well-understood for the Euclidean norm.

Subspace Embeddings. In this problem, one would like a distribution on linear maps $\mathbf{S} \in \mathbb{R}^{r \times n}$, for which with constant probability over the choice of $\mathbf{S}$, for any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_1$. Note that $\mathbf{S}$ preserves the lengths of an infinite number of vectors, namely, the entire column span of $\mathbf{A}$. Subspace embeddings arise in least absolute deviation regression Sohler and Woodruff (2011); Clarkson et al. (2013); Meng and Mahoney (2013); Woodruff and Zhang (2013); Clarkson and Woodruff (2017) and entrywise ℓ_1-low rank approximation Song et al. (2017); Ban et al. (2019); Mahankali and Woodruff (2021), among other places.

1. Their bounds are stated for subspaces, but when applied to n arbitrary points result in this as both an upper and a lower bound, with differing polynomial factors in the exponent. We give more details in Remark A.6 of Section A.

Since a subspace embedding maps the entire subspace into a lower dimensional subspace of ℓ_1, one can impose arbitrary constraints on $\mathbf{x}$, e.g., non-negativity, manifold constraints, regularization, and so on, after computing $\mathbf{SA}$. The resulting problem in the sketch space is convex if the constraints are convex.

For the analogous problem in the Euclidean norm, there is a linear oblivious sketching matrix S with $O((d + \log(1/\delta))/\varepsilon^2)$ rows, which is best possible Clarkson and Woodruff (2009); Nelson and Nguyen (2014); Woodruff (2014).

For the 1-norm, we understand much less. The best upper bound Wang and Woodruff (2019) for an oblivious subspace embedding is for constant ε and δ and gives a sketching dimension of $2^{2^{O(d)}}$. This bound is obtained by instantiating the $2^{\text{poly}(n)}$ bound above with $n = d^{O(d)}$, and union bounding over the points in a net of a subspace. The lower bound on the sketching dimension is, however, only $2^{\Omega(\sqrt{d})}$, representing an exponential gap in our understanding for this fundamental problem Wang and Woodruff (2019).

Independence Testing. Another important problem using dimensionality reduction for ℓ_1 is testing independence in a stream. This problem was introduced by Indyk and McGregor Indyk and McGregor (2008) and is the following: letting $[d] = \{1, 2, \ldots, d\}$, suppose you are given a stream of items $(i_1, \ldots, i_q) \in [d]^q$. These define an empirical joint distribution P on the q modes defined as follows: if $f(i_1, ..., i_q)$ is the number of occurrences of $(i_1, \ldots, i_q)$ in a stream of length m, then $P(i_1, \ldots, i_q) = \frac{1}{m} f(i_1, \ldots, i_q)$. One can also define the marginal distributions P_j, for $j = 1, 2, \ldots, q$, where for $i \in [d]$ we have $P_j(i) = \frac{1}{m} \sum_{i_1,\ldots,i_{j-1},i_{j+1},\ldots i_q} f(i_1, \ldots, i_{j-1}, i, i_{j+1}, \ldots i_q)$. The goal is to compute $\|P - Q\|_1$ with $Q = P_1 \otimes P_2 \otimes \cdots \otimes P_q$, that is, the 1-norm of the difference of the joint distribution and the product of marginals. If the q modes were independent, then this difference would be 0, as P would be a product distribution. In general this measures the distance to independence.

It is important to note that if one were given $P_1, \ldots, P_q$ and P, then one could explicitly compute $P_1 \otimes P_2 \otimes \cdots \otimes P_q$, and then compute the median-based sketch $\mathbf{S}(P - P_1 \otimes P_2 \otimes \cdots \otimes P_q)$ of Indyk Indyk (2006b) above. The issue is that in the data stream model, the vectors $P, P_1, \ldots, P_q$ are too large to store, and while it is easy to update $\mathbf{S}(P)$ given a new tuple in the stream (namely, $\mathbf{S}(P) \leftarrow \mathbf{S}(P) + \mathbf{S}_{i_1,\ldots,i_q}$, where $\mathbf{S}_{i_1,\ldots,i_q}$ is the column of $\mathbf{S}$ indexed by the new stream element $(i_1, \ldots, i_q)$), it is not clear how to update $\mathbf{S}(P_1 \otimes P_2 \otimes \cdots \otimes P_q)$ in a stream. Consequently, a natural approach is to maintain sketches $\mathbf{S}^1 P_1, \mathbf{S}^2 P_2, \ldots, \mathbf{S}^q P_q$ as well as $\mathbf{S}P$, and combine these at the end of the stream. A natural way to combine them is to let $\mathbf{S} = (\mathbf{S}^1) \otimes (\mathbf{S}^2) \otimes \cdots \otimes (\mathbf{S}^q)$ be the tensor product of the sketches on each mode.

For the corresponding problem of estimating the Euclidean distance $\|P - P_1 \otimes P_2 \otimes \cdots \otimes P_q\|_2$, recent work Ahle et al. (2020) implies that this can be done with a very small sketching dimension of $O(n/\varepsilon^2)$, though such work makes use of the Johnson Lindenstrauss lemma and completely fails for the 1-norm.

Despite a number of works on independence testing for the 1-norm in a stream Indyk and McGregor (2008); Braverman et al. (2010); Braverman and Ostrovsky (2010a); McGregor and Vu (2015), the best upper bound is due to Braverman and Ostrovsky Braverman and Ostrovsky (2010a) with a sketching dimension of $(\varepsilon^{-1} \log d)^{q^{O(q)}}$, which, while logarithmic in d, is doubly exponential in q. A natural question is whether this can be improved.

1.1. Our Results

We give exponential improvements in the sketching dimension of linear oblivious maps for both ℓ_1-subspace embeddings and ℓ_1-independence testing.

Subspace Embeddings: We design a distribution over random matrices $\mathbf{S} \in \mathbb{R}^{r\times n}$, where $r = 2^{\text{poly}(d/(\varepsilon\delta))}$, so that given any matrix $A \in \mathbb{R}^{n\times d}$, with probability at least $1-\delta$, simultaneously for all $\mathbf{x}$, $\|\mathbf{SAx}\|_1 = (1\pm\varepsilon)\|\mathbf{Ax}\|_1$. We present both a sparse embedding which has a dependence of $\log n$ in the base of the exponent, as well as a dense embedding which removes this dependence on n entirely.

Theorem 1.1 (Sparse embedding, restatement of Theorem B.1) *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Then there exists a sparse oblivious ℓ_1 subspace embedding $\mathbf{S}$ into k dimensions with*

$$k = \text{poly}(d, \varepsilon^{-1}, \delta^{-1}, \log n)^{d/\delta\varepsilon}$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\{(1-\varepsilon)\|\mathbf{Ax}\|_1 \leqslant \|\mathbf{SAx}\|_1 \leqslant (1+\varepsilon)\|\mathbf{Ax}\|_1\} \geqslant 1-\delta.$$

Corollary 1.2 (Dense embedding, restatement of Corollary B.2) *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Then there exists an oblivious ℓ_1 subspace embedding $\mathbf{S}$ into k dimensions with*

$$k = \exp\Big(\tilde{O}(d^2/\delta\varepsilon^3)\Big)$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\{(1-\varepsilon)\|\mathbf{Ax}\|_1 \leqslant \|\mathbf{SAx}\|_1 \leqslant (1+\varepsilon)\|\mathbf{Ax}\|_1\} \geqslant 1-\delta.$$

This is an exponential improvement over the previous bound of $r = 2^{2^{\Omega(d)}}$ Wang and Woodruff (2019), which held for constant ε and δ. Our bound is optimal, up to a polynomial factor in the exponent, given the $2^{\text{poly}(d)}$ lower bound for constant ε and δ Wang and Woodruff (2019). An important feature of $\mathbf{S}$ is that $\mathbf{S}\cdot\mathbf{A}$ can be computed in an expected $O(\text{nnz}(\mathbf{A}))$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries of $\mathbf{A}$. This is in contrast to the embedding of Wang and Woodruff (2019), which requires $2^{2^{\Omega(d)}}\cdot\text{nnz}(\mathbf{A})$ time.

Independence Testing: We design a distribution over matrices $\mathbf{S} \in \mathbb{R}^{k\times n}$, where $k = \text{poly}(q\varepsilon^{-1}\log d)$, so that given any q-mode tensor $\mathbf{A} \in (\mathbb{R}^d)^{\otimes q}$, one can estimate the entrywise 1-norm $\|\mathbf{A}\|_1$ from $\mathbf{S}(\mathbf{A})$. Moreover, $\mathbf{S} = \mathbf{T}^{\otimes q}$ and so given vectors $\mathbf{u}_1,\ldots,\mathbf{u}_q \in \mathbb{R}^d$, one can compute $\mathbf{S}(\mathbf{u}_1\otimes\mathbf{u}_2\otimes\cdots\otimes\mathbf{u}_q)$ in time $2^{O(q^2)}(\varepsilon^{-1}q\log d)^{O(q)}$, which is much faster than the d^q time required to form $\mathbf{u}_1\otimes\mathbf{u}_2\otimes\cdots\otimes\mathbf{u}_q$. Our linear map can be applied in a stream since we can sketch each marginal and then take the tensor product of sketches, yielding a streaming algorithm for independence testing using $2^{O(q^2)}(\varepsilon^{-1}q\log d)^{O(q)}$ bits of space.

Theorem 1.3 (Restatement of Theorem 5) *Suppose that the stream length $m = \text{poly}(d^q)$. There is a randomized sketching algorithm which outputs a $(1\pm\varepsilon)$-approximation to $\|P-Q\|_1$ with probability at least 0.9, using $\exp(O(q^2 + q\log(q/\varepsilon) + q\log\log d))$ bits of space. The update time is $\exp(O(q^2 + q\log(q/\varepsilon) + q\log\log d))$.*

This improves the previous doubly exponential $(\varepsilon^{-1}\log d)^{q^{O(q)}}$ space bound Braverman and Ostrovsky (2010a).

For subspace embeddings, we also study the setting when $\mathbf{A}$ is itself drawn from distributions with certain properties, and obtain a polynomial embedding dimension. This captures natural statistical problems when the design matrix $\mathbf{A}$ for regression, is itself random. Our various results here are discussed in Section E.

A byproduct of our sketch is the ability to preserve the 1-norm of a matrix $\mathbf{P}$ by left and right multiplying by independent draws $\mathbf{S}^1$ and $\mathbf{S}^2$ of our sketch, where we show that $\Theta(\|\mathbf{P}\|_1) \leqslant \|\mathbf{S}^1\mathbf{P}\mathbf{S}^2\|_1 = O(1/\alpha^2)\|\mathbf{P}\|_1$ where $\mathbf{S}^1\mathbf{P}\mathbf{S}^2$ is a $d^\alpha \times d^\alpha$ matrix. Here $\alpha \in (0,1)$ can be any constant; previously, no such trade-off was known.

Theorem 1.4 (Restatement of Theorem C.2) *Let $\delta \in (0,1)$ and $\alpha \in (0,1)$. Then there exists a sparse oblivious ℓ_1 entrywise embedding $\mathbf{S}$ into k dimensions with*

$$k = \left(\frac{d}{\delta}\log n\right)^{\alpha} \mathrm{poly}(\delta^{-1}, \log n)$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\left\{\Omega(1)\|\mathbf{A}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\|_1 \leqslant O\left(\frac{1}{\delta\alpha}\right)\|\mathbf{A}\|_1\right\} \geqslant 1-\delta.$$

We also give a matching lower bound showing that for any oblivious sketch $\mathbf{S}^1$ with r rows, the distortion between $\|\mathbf{S}^1\mathbf{P}\|_1$ and $\|\mathbf{P}\|_1$ is $\Omega\left(\frac{\log d}{\log r}\right)$. Thus, with $r = d^\alpha$ dimensions, the distortion must be at least

$$\frac{\log d}{\log r} = \frac{\log d}{\log d^\alpha} = \frac{1}{\alpha}.$$

Theorem 1.5 (Restatement of Theorem C.8) *Let $\mathbf{S}$ be a fixed $r \times d$ matrix. Then there is a distribution μ over $d \times d$ matrices such that if*

$$\mathbf{Pr}_{\mathbf{A}\sim\mu}\left(\|\mathbf{A}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\|_1 \leqslant \kappa\|\mathbf{A}\|_1\right) \geqslant \frac{2}{3}$$

then $\kappa = \Omega((\log d)/(\log r))$.

For independence testing, we also give algorithms for any distance measure with a polylogarithmic-sized sketch and satisfying an approximate triangle inequality; these include many functions in Braverman and Ostrovsky (2010b). For example, we handle the robust Huber loss and ℓ_p-measures for $0 < p < 2$.

1.2. Our Techniques

We begin by explaining our techniques for subspace embeddings, and then transition to independence testing.

1.2.1. SUBSPACE EMBEDDINGS

The linear oblivious sketch we use is a twist, both algorithmically and analytically, to a methodology originating from the data stream literature for approximating frequency moments Indyk and Woodruff (2005); Bhuvanagiri et al. (2006). These methods involve sketches which subsample the coordinates of a vector at geometrically decreasing rates $1, 1/2, 1/4, 1/8, \ldots, 1/n$, and apply an independent CountSketch matrix Charikar et al. (2002) (see Definition A.1) to the surviving coordinates at each scale. Analyses of this sketch for data streams does not apply here, since it involves nonlinear median operations, but here we must embed ℓ_1 into ℓ_1. These sketches *have* been used for embedding single vectors or matrices in ℓ_1 into ℓ_1, called the *Rademacher sketch* in Verbin and Zhang (2012), and the M*-sketch* in Clarkson and Woodruff (2015). However the approximation guarantees in these works are significantly worse than what we achieve, and we improve them by (1) changing the actual sketch to "randomized boundaries" and (2) changing the analysis of the sketch to track the behavior of the ℓ_1-leverage score vector, which captures the entire subspace, and tracking it via a new mix of expected and high probability events.

We now explain these ideas in more detail. To motivate our sketch, we first explain the pitfalls of previous sketches.

Cauchy Sketches Sohler and Woodruff (2011); Wang and Woodruff (2019). The previous best $O(1)$ distortion ℓ_1 oblivious subspace embedding of Wang and Woodruff (2019), which achieved a sketching dimension of $2^{2^{O(d)}}$, was based on analyzing a sketch $\mathbf{S}$ of i.i.d. Cauchy random variables. The only analyses of such random variables we are aware of, in the context of subspace embeddings, works by truncating the random variables so that they have a finite expectation, and then analyzing the behavior of the random variable $\|\mathbf{Sy}\|_1$, for an input vector $\mathbf{y}$ in expectation. It turns out that the expectation of this random variable can be much larger than the value it takes with constant probability, as it is very heavy-tailed. Namely, the expected value of $\|\mathbf{Sy}\|_1$ after truncation is $\Theta(\log n)\|\mathbf{y}\|_1$, which makes it unsuitable for the sketching dimension that we seek.

Rademacher and M Sketches Verbin and Zhang (2012); Clarkson and Woodruff (2015). Using techniques from the data stream literature, the *Rademacher sketch* of Verbin and Zhang (2012) and the M*-sketch* of Clarkson and Woodruff (2015) achieve an $O(1)$-approximation for a single vector by subsampling rows of $\mathbf{y}$ with probability p and rescaling by $1/p$ at $O(\log n)$ scales $p = 1, 1/2, 1/4, 1/8, \ldots, 1/n$. This approach allows us to more finely track the random variables in our sketch, and serves as the starting point of our sketch. Note that for a single scale p and a single coordinate $\mathbf{y}_i$, the expected contribution of the subsampled and rescaled coordinate is

$$\frac{1}{p} \cdot p \cdot |\mathbf{y}_i| = |\mathbf{y}_i|.$$

Then in expectation, the $O(\log n)$ subsampling levels give a $O(\log n)$ factor approximation, which is the same as that of a Cauchy sketch. However, due to the geometrically decreasing sampling rates, we are able to argue that with good probability the coordinate does not survive more than $O(1)$ levels. Thus we effectively "beat the expectation", showing that the random variable is much less than what its expectation would predict, with good probability. We illustrate this with an example.

Suppose the first $\sqrt{n}$ coordinates of $\mathbf{y}$ equal $\frac{1}{\sqrt{n}}$, and remaining $n - \sqrt{n}$ coordinates equal $\frac{1}{n}$. Then $\|\mathbf{y}\|_1 = 2(1 - o(1))$. If we subsample at geometric rates $1, 1/2, 1/4, \ldots, 1/n$ and use $t = O(1)$ hash buckets in CountSketch in each scale, then for rates larger than $1/\sqrt{n}$, the random

signs in each CountSketch bucket cancel out and the absolute value of the bucket concentrates to its Euclidean norm, which is much smaller than its 1-norm. At the rate $p = 1/\sqrt{n}$, we expect a single survivor from the first $\sqrt{n}$ coordinates of $\mathbf{y}$. We call this the *ideal rate* for the first $\sqrt{n}$ coordinates of $\mathbf{y}$. There are also about $\sqrt{n}$ survivors from the remaining $n - \sqrt{n}$ coordinates of $\mathbf{y}$ at this ideal rate, but these $\sqrt{n}$ survivors concentrate to their Euclidean norm in each CountSketch bucket, which will be about $1/n^{3/4}$, and negligible compared to the value $1/\sqrt{n}$. This lone survivor will be scaled up by $\sqrt{n}$, giving a contribution of 1 to the overall 1-norm. Similarly, at the subsampling rate of $1/n$, we expect one surviving coordinate of $\mathbf{y}$, it is scaled up by n, and it gives an additional contribution of about 1 to the overall 1-norm. Overall, this gives a good approximation to $\|\mathbf{y}\|_1$, which is $2(1 - o(1))$.

While the above gives a good approximation, the expected value of the 1-norm of $\mathbf{Sy}$ is a much larger $\Theta(\log n)$. Indeed, consider subsampling rates $1/(2\sqrt{n}), 1/(4\sqrt{n}), 1/(8\sqrt{n}), \ldots$. For each of these, the single survivor of the first $\sqrt{n}$ coordinates of y has probability $1/2, 1/4, 1/8, \ldots$, of surviving each successive level. If it survives, it is scaled up by $2, 4, 8, \ldots$, giving an overall expectation of $\Theta(\log n)$. Thus, the expectation is not what we should be looking at, but rather we should be conditioning on the event that no items among the first $\sqrt{n}$ surviving beyond the rate $1/\sqrt{n}$.

Ingredient 1: Aggressive Subsampling and Randomized Boundaries. So far, this is standard. Indeed, the *Rademacher sketch* in Verbin and Zhang (2012) and the *M-Sketch* in Clarkson and Woodruff (2015) achieve an $O(1)$-approximation for a single vector and argue this way. But these works cannot achieve $(1+\varepsilon)$-approximation with good probability, since it is already problematic if the single survivor of the first $\sqrt{n}$ coordinates of $\mathbf{y}$ survives one additional subsampling rate beyond its ideal rate, and this happens with constant probability. This motivates our first fix: instead of subsampling at rates $1/2^i$, for $i = 0, 1, 2, \ldots, O(\log n)$, we subsample at a much more aggressive $\exp(\varepsilon^{-1}\operatorname{polylog}(n))^i$ for $i = 0, 1, 2, \ldots, O(\log n)$, and furthermore, randomly shift these subsampling rates as well.

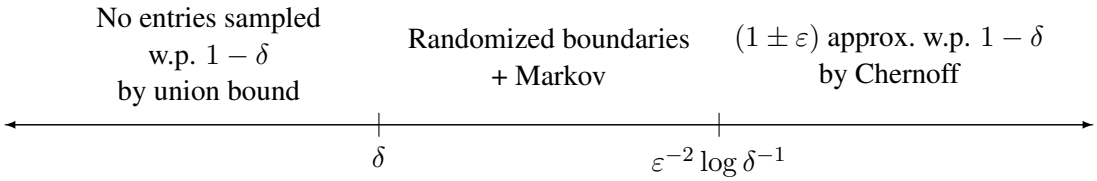

Figure 1: Casework on pm

To see why this is a good idea, consider a level set of weight w, which is the multiset of coordinates of $\mathbf{y}$ with absolute value $\Theta(w)$ (think of w as $[2^j, 2^{j+1})$ for some j) that is subsampled at rate p and rescaled by $1/p$. Let the size of the level set be m. We case on pm (see Figure 1). If $pm \geqslant \varepsilon^{-2} \log \frac{1}{\delta}$, then Chernoff bounds imply that this concentrates to the expected mass of pm with probability at least $1 - \delta$. On the other hand, if $pm < \delta$, then by a union bound, there is a δ probability that any of the m elements in the level set are sampled. By taking $\delta = 1/\log^2 n$, we see that by a union bound over the at most $\log n$ level sets and $\log n$ sampling rates p, any level set with size m and subsampling rate p with $pm \notin [\delta, \varepsilon^{-2} \log \delta^{-1}]$ either samples $(1 \pm \varepsilon)$ of the expected mass, or doesn't sample the level at all, with constant probability. Then, for these levels, our earlier analyses involving CountSketch apply and in fact give us a $(1 \pm O(\varepsilon))$ approximation.

However, for the level sets and the sampling rates with $pm \in [\delta, \varepsilon^{-2}\log\delta^{-1}]$, we cannot make any meaningful statements about these levels with high accuracy and probability. To remedy this situation, we randomize our choice of the sampling rates p themselves and bound the contribution from these levels with a Markov expectation bound. To this end, we let $W = \varepsilon^{-2}\delta^{-1}\log\delta^{-1}$ be the size of this bad window, we let $B = \exp(\varepsilon^{-1}\log W)$ be our branching factor, and we choose our sampling rates to be $p_i = B^{-u}B^{-i}$ for a uniformly random $u \sim [0,1]$. Note then that the probability that a given sampling level p_i falls in the window $p_i m \in [\delta, \varepsilon^{-2}\log\delta^{-1}]$ is at most ε, since after taking logarithms, the bad window is an ε fraction of the range of the uniformly random shift u. Now note that for each level set of size m and weight w, there are only $O(1)$ sampling levels p_i that have a nonzero probability such that $p_i m \in [\delta, \varepsilon^{-2}\log\delta^{-1}]$, and these levels contribute an expected $\varepsilon \cdot p \cdot p^{-1} \cdot m \cdot w = \varepsilon m w$ amount of ℓ_1 mass, so summing over all level sets, the *expected* contribution from these bad sampling rates is at most an ε fraction of the total ℓ_1 mass $\|\mathbf{y}\|_1$.

This is an example of how subsampling gives us more flexibility than sketches using Cauchy random variables - even though the expectation is large, we can argue with arbitrarily large constant probability we obtain a $(1+O(\varepsilon))$-approximation by separating the analysis into an expectation for some levels and a union bound for others. One also needs to argue that no vector has its 1-norm shrink by more than a $(1-\varepsilon)$-factor, which is simpler and similar to previous work Clarkson and Woodruff (2015). Here the idea is that for every level set of coordinates of $\mathbf{y}$, by Chernoff bounds, there are enough survivors in a level set at its ideal rate and that the noise in CountSketch buckets will be small. Our analysis so far is novel, and we note that prior analyses of subsampling Verbin and Zhang (2012); Clarkson and Woodruff (2015) could not obtain a $1+O(\varepsilon)$-approximation even for a fixed vector.

However, we are still in trouble - the above analysis gives a $(1+O(\varepsilon))$-approximation, but only a constant probability of success due to the Markov bound applied to the bad sampling rates. We could more aggressively subsample, namely, at rate roughly $1/2^{2^{O(d)}}$ and with $2^{2^{O(d)}}$ buckets, and then we could make the failure probability $(\varepsilon)^{O(d)}$ for a fixed vector, which is now small enough to union bound over an ε-net of vectors in a d-dimensional subspace. This is enough to recover the same sketching dimension as the sketch in Wang and Woodruff (2019), which instead consisted of an $r \times n$ matrix of i.i.d. Cauchy random variables. There it was shown that with probability $1 - O\left(\frac{\log\log r}{\log r}\right)$, for any fixed vector $\mathbf{y}$, $\|\mathbf{S}\mathbf{y}\|_1 = \Theta(1)\|\mathbf{y}\|_1$. The idea was then to take a union bound over $2^{O(d)}$ vectors in a net for the subspace, which constrains $\frac{\log\log r}{\log r} \leqslant 2^{-\Theta(d)}$, resulting in an $r = 2^{2^{O(d)}}$ overall dependence. With minor modifications, one can achieve $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_1$ for all $\mathbf{x}$ by setting $r = 2^{2^{O(d/\varepsilon^2)}}$. This is the best one can achieve for an arbitrary set of $2^{O(d)}$ vectors, as can be deduced from the lower bound in Wang and Woodruff (2019); see Section A for details.

Ingredient 2: ℓ_1 Leverage Scores. One might suspect that the above approach is optimal, since union bounding over $2^{O(d)}$ arbitrary points does give an optimal sketching dimension for subspace embeddings for the Euclidean norm. It turns out though that for the 1-norm this is not the case, and one can do exponentially better by using the fact that these $2^{O(d)}$ points all live in the same d-dimensional subspace. Indeed, instead of making a net argument, our analysis proceeds through the *ℓ_1-leverage score vector* (see Definition B.11), which provides a nonuniform importance sampling distribution that is analogous to the standard leverage scores for ℓ_2.

With these ℓ_1 leverage scores in hand, we proceed as discussed previously, choosing a uniformly random shift $u \in [0,1]$ and subsampling at rates $1/((\log n)^{\text{poly}((d/\varepsilon)(i+u))})$ for $i = 0, 1, 2, \ldots, O(\log n)$,

and also increasing our number of CountSketch buckets in each subsampling level to $(\log n)^{\mathrm{poly}(d/\varepsilon)}$. Now we can show that the expected ℓ_1-norm of the ℓ_1 leverage score vector $\boldsymbol{\lambda}$ that survives an additional level is only $\varepsilon\|\boldsymbol{\lambda}\|_1/d$. Noting that $\|\boldsymbol{\lambda}\|_1 = d$, this bound is $O(\varepsilon)$ with constant probability. But the entries of $\boldsymbol{\lambda}$ uniformly bound the corresponding entries of any vector $\mathbf{y}$ in the subspace with $\|\mathbf{y}\|_1 = 1$, and thus we obtain that for all vectors in the subspace, the total expected ℓ_1-contribution from level sets that are one subsampling rate beyond their ideal rate is $O(\varepsilon)\|\mathbf{y}\|_1$. Since the subsampling rate is $(\log n)^{-\mathrm{poly}(d/\varepsilon)}$, the expected number of survivors two or more levels out is small enough to union bound over all net vectors. Finally, to remove the $\log n$ factor in our sketch, making it independent of the original dimension n, we can compose our embedding with the $2^{2^{O(d)}}$ ℓ_1 oblivious subspace embedding of Wang and Woodruff (2019); we are able to adapt their $O(1)$-approximation to achieve a $(1+\varepsilon)$-approximation with $2^{2^{O(d/\varepsilon^2)}}$ dimensions, and consequently in our sketch, $\log n = 2^{O(d/\varepsilon^2)}$. Our full discussion is in Section B.

1.2.2. A TRANSITION TO TENSORS

One could hope to use our techniques for subspaces to obtain sketches for the sum of q-mode tensors, which could then be used for independence testing in a stream. Consider the simple example of a 2-mode tensor, i.e., a $d \times d$ matrix $\mathbf{P}$. As described above, a streaming-amenable way of sketching this would be to find a sketch $\mathbf{S} : \mathbb{R}^{d^2} \to \mathbb{R}^{k^2}$ of the form $\mathbf{S} = \mathbf{S}^1 \otimes \mathbf{S}^2$, where $\mathbf{S}^1, \mathbf{S}^2$ are maps from $\mathbb{R}^d$ to $\mathbb{R}^k$. In this case, we have that $\mathbf{S}(\mathbf{P}) = \mathbf{S}^1 \cdot \mathbf{P} \cdot (\mathbf{S}^2)^\top$, where $\cdot$ denotes matrix multiplication.

One aspect of our sketch above is that we can achieve a tradeoff: instead of looking at one subsampling rate beyond the ideal rate for a given level set of a vector, we can look at $1/\alpha$ rates for $\alpha \in (0,1)$. Then if we look at $\|\mathbf{S}\mathbf{y}\|_1$ for a column $\mathbf{y}$ of $\mathbf{P}$, its expected cost for these $1/\alpha$ rates is $O(1/\alpha)\|\mathbf{S}\mathbf{y}\|_1$. If we use roughly $(d\log d)^\alpha$ buckets in each CountSketch, together with subsampling rate roughly $(d\log d)^{-\alpha}$, then after $O(1/\alpha)$ rates beyond the ideal rate for a given level set of a vector, the probability the level set survives is at most $\left(\frac{1}{(d\log d)^\alpha}\right)^{O(1/\alpha)} \ll O\left(\frac{1}{d\log d}\right)$, which is so small that we can union bound over all columns of $\mathbf{P}$ and all level sets in each column. Consequently, we can condition on this event, and take an expectation over the $O(1/\alpha)$ rates nearest to the ideal rate of each level set in each column to obtain an overall $O(1/\alpha)$ approximation with roughly $(d\log d)^\alpha$ memory. One can also show that with constant probability, the 1-norm does not decrease by more than a constant factor, and thus, with constant overall probability, $\Omega(\|\mathbf{P}\|_1) \leqslant \|\mathbf{S}^1\mathbf{P}\|_1 = O(1/\alpha)\|\mathbf{P}\|_1$. Applying $\mathbf{S}^2$ to the matrix $\mathbf{S}^1\mathbf{P}$ we can conclude that with constant overall probability, $\Omega(\|\mathbf{P}\|_1) \leqslant \|\mathbf{S}^1\mathbf{P}\mathbf{S}^2\|_1 = O(1/\alpha^2)\|\mathbf{P}\|_1$. Our overall sketching dimension is $d^{2\alpha} \ll d$ if $\alpha \ll 1$. Thus, the memory we achieve is a significant improvement over the trivial d^2 bound, our sketch $\mathbf{S} = \mathbf{S}^1 \otimes \mathbf{S}^2$ is a tensor product, and we achieve an $O(1/\alpha^2)$-approximation. Ours is the first sketch to achieve a tradeoff, as the Rademacher sketch of Verbin and Zhang (2012) does not apply in this case[2].

Unfortunately, if we want constant distortion, our single-mode sketch size k will be $d^{2\alpha}$, which means for constant α, it is not strong enough to obtain a polylogarithmic dependence on d. In fact, we show that for any $d \times d$ matrix $\mathbf{P}$, if you compute $\mathbf{S}\mathbf{P}$ for an oblivious sketch $\mathbf{S}$ with t rows, the estimator $\|\mathbf{S}\mathbf{P}\|_1$ is at best an $O\left(\frac{\log d}{\log t}\right)$-approximation to $\|\mathbf{P}\|_1$. Indeed, one can show this

2. The notion of the Rademacher dimension in Verbin and Zhang (2012) is at least $\sqrt{d}$, and their sketch size is at least the Rademacher dimension to the 5-th power.

already for the distribution in which with probability $1/2$, $\mathbf{P} \in \mathbb{R}^{d\times d}$ is an i.i.d. Cauchy matrix, and with probability $1/2$, $\mathbf{P}$ has its first t columns being i.i.d. Cauchy random variables, scaled by d/t, and remaining columns equal to 0. In both cases $\|\mathbf{P}\|_1 = \Theta(d^2 \log d)$, but in the first case $\|\mathbf{SP}\|_1 = O(d \log t \|\mathbf{S}\|_1)$, while in the second case $\|\mathbf{SP}\|_1 = \Omega(d \log d \|\mathbf{S}\|_1)$, both with constant probability. These algorithms and lower bounds are discussed in Section C.

Fortunately, for independence testing, we only need to approximate the 1-norm of a single tensor, and so our estimator can be a non-convex median-based estimator, which we now show how to utilize.

1.2.3. INDEPENDENCE TESTING

Our sketch $\mathbf{S} = \mathbf{S}^1 \otimes \mathbf{S}^2 \otimes \cdots \otimes \mathbf{S}^q$ is a tensor product of q sketches, each itself being a sketch for estimating the 1-norm of a d-dimensional vector with a $\log(1/\delta)$ dependence. We must choose the $\mathbf{S}^i$ carefully, and cannot take the $\mathbf{S}^i$ to be an arbitrary black box sketch for estimating the 1-norm, even with a non-linear high probability estimator. As an illustration, suppose $q = 2$ and we have a $d \times d$ matrix $\mathbf{P}$ and we compute $\mathbf{S}^1\mathbf{P}\mathbf{S}^2$, where $\mathbf{S}^1$ and $\mathbf{S}^2$ are i.i.d. Cauchy matrices with $r = O(\varepsilon^{-1} \log d)$ small dimension with corresponding median of absolute values estimator, i.e., the sketch of Indyk (2006a) above. Then, applying the estimator of $\mathbf{S}^2$ to each row of $\mathbf{S}^1\mathbf{P}$, we would have that our overall estimate is $(1 \pm \varepsilon)\|\mathbf{S}^1\mathbf{P}\|_1$ with probability $1 - 1/\operatorname{poly}(d)$. The issue is that, for constant ε, if $\mathbf{P} = (1, 1, 1, \ldots, 1) \otimes (1, 0, 0, \ldots, 0)$, then $\|\mathbf{S}^1\mathbf{P}\|_1 = \Theta(d \log r)$ with large probability, while if $\mathbf{P} = \mathbf{I}_d$, the $d \times d$ identity matrix, then $\|\mathbf{S}^1\mathbf{P}\|_1 = \Theta(d \log d)$ with large probability. To see this, if $\mathbf{P} = (1, 1, 1, \ldots, 1) \otimes (1, 0, 0, \ldots, 0)$, note that the i-th row of $\mathbf{S}^1\mathbf{P} = d \cdot (C^i, 0, \ldots, 0)$, where C^i is a standard Cauchy, and the $C^1, \ldots, C^r$ are independent. About a $\Theta(2^{-j})$ fraction of the $|C^i|$ will be 2^j, and so with constant probability $\|\mathbf{S}^1\mathbf{P}\|_1 = \Theta(d \log r)$. On the other hand, if $\mathbf{P} = \mathbf{I}_d$, then $\mathbf{S}^1\mathbf{P} = \mathbf{S}^1$, which is an $r \times d$ matrix of i.i.d. Cauchy random variables, and the same reasoning shows with constant probability that $\|\mathbf{S}^1\mathbf{P}\|_1 = \Theta(d \log(rd))$, which is almost a $\log d$ factor larger than the other case. Thus, we cannot decode mode by mode with a generic high probability sketch for the 1-norm.

Perhaps surprisingly, we show that a different choice of $\mathbf{S}^i$, which is itself an existing sketch for estimating the 1-norm of a d-dimensional vector with a $\log(1/\delta)$ dependence, *does work*. In more detail, the sketch of Indyk and Woodruff (2005) works by defining level sets of coordinates of $\mathbf{x}$ according to their magnitudes and subsamples the coordinates at different rates. For each level set, if it contributes a non-negligible fraction to $\|\mathbf{x}\|_1$, there is a subsampling level for which (1) there are sufficiently many survivors from the level set in this subsampling level and (2) these survivors are so-called ℓ_2-heavy hitters (see, e.g., Charikar et al. (2002)) among all the survivors in this subsampling level. Hence, recovering the heavy hitters at each subsampling rate allows us to estimate the contribution of each level set to $\|\mathbf{x}\|_1$. Here a median is used when applying CountSketch to ensure that we succeed with high probability. This single mode sketch has been applied to ℓ_1-estimation in various places Andoni et al. (2009); Levin et al. (2018). We refer to this as a *SubsamplingHeavyHitters sketch* in the following discussion.

Our overall sketch $\mathbf{S} = \mathbf{S}^1 \otimes \mathbf{S}^2 \otimes \cdots \otimes \mathbf{S}^q$, where each $\mathbf{S}^i$ is a SubsamplingHeavyHitters sketch. Moreover, $\mathbf{S} = \mathbf{S}^1 \otimes \cdots \otimes \mathbf{S}^q$, and so given vectors $P^1, \ldots, P^q \in \mathbb{R}^d$ in a stream, one can maintain $\mathbf{S}^i P^i$ for $i = 1, \ldots, q$, as well as $\mathbf{S}P$ for any vector $P \in \mathbb{R}^{d^q}$. In particular, in the context of independence testing, the P^i could be the empirical marginal distributions and P the empirical joint distribution. We show that $\mathbf{S}$ can be used to estimate the ℓ_1-norm of an underlying arbitrary vector $x \in \mathbb{R}^{d^q}$ (which will be taken to be $P - P^1 \otimes \cdots \otimes P^q$). We do this by viewing $\mathbf{S}^q$ as being

applied to each row of a flattened $t^{q-1} \times d$ matrix, where t is the common sketching dimension of the $\mathbf{S}^i$. This matrix is defined as follows. We flatten x to a $d^{q-1} \times d$ matrix X. We then consider the "partially sketched" $d^{q-1} \times d$ matrix, where the i-th column is $\mathbf{S}^1 \otimes \mathbf{S}^2 \otimes \cdots \otimes \mathbf{S}^{q-1}$ applied to the i-th column $X_{*,i}$ of X. This gives us a $t^{q-1} \times d$ matrix Y, and this is the matrix whose rows we apply $\mathbf{S}^q$ to. Now $\mathbf{S}^q$ is a SubsamplingHeavyHitters sketch, but instead of having a signed sum of single coordinates in each CountSketch bucket, we have a signed sum of columns of Y in each bucket, which are themselves sketches of d^{q-1}-dimensional vectors, where the sketching matrix is itself a tensor product of smaller sketching matrices.

The problem is that $\mathbf{S}^q$ estimates the number of columns of a matrix in a level set (here the level sets are groups of columns with approximately the same 1-norm) by hashing columns together and estimating the size of each level set, where columns are in the same level set if they have approximately the same 1-norm. Fortunately, since $\mathbf{S}^1 \otimes \cdots \otimes \mathbf{S}^{q-1}$ is still a linear map, hashing the sketched columns (sketched by $\mathbf{S}^1 \otimes \cdots \otimes \mathbf{S}^{q-1}$) together is the same as taking the sketch (by $\mathbf{S}^1 \otimes \cdots \otimes \mathbf{S}^{q-1}$) of the hashed columns together. However, it is still unclear what the 1-norm of the sketch of the hashed columns is. In fact, it cannot be concentrated with high probability by the above discussion. Fortunately, for each bucket in a CountSketch associated with a subsampling rate in $\mathbf{S}^q$, we can use our knowledge of $\mathbf{S}^1 \otimes \cdots \otimes \mathbf{S}^{q-1}$ to *recursively estimate* the 1-norm inside of that bucket. This recursive estimation involves applying $\mathbf{S}^{q-1}$ to the rows of a $t^{q-2} \times d$ matrix Z, computing recursive estimates, and so on. Finally, we use these recursive estimates to estimate the level sets of columns of the matrix X, and ultimately build and output the estimator provided by $\mathbf{S}^q$.

The main issue we still face is how to handle the blowup in approximation ratio and error probability in each recursive step. In each $\mathbf{S}^i$ we would like to randomize boundaries to avoid overcounting when estimating level set sizes in the estimator. However, the approximation error grows as we decode more modes. The most natural approach, if the error after decoding the i-th mode is $(1+\eta)$, is to randomize boundaries so that the probability is $O(\eta)$ of landing near a boundary, and consequently not being included in the estimator, when decoding $\mathbf{S}^{i+1}$. However, this blows up the approximation to $(1+\eta)^2$. Unfolding the recursion, we get a $(1+\varepsilon)^{\tilde{O}(2^q)}$ overall approximation. Setting our initial ε to $\varepsilon/2^{\tilde{O}(q)}$, we can make the overall approximation $1+\varepsilon$. This yields a $2^{O(q)}$ factor in the sketching dimension on each mode and thus a $2^{O(q^2)}$ factor in the sketching dimension in the overall tensor product.

It seems difficult to improve the $2^{O(q^2)}$ bound. To improve this bound, we need to make the error smaller than $(1+\eta)^2$ in the $(i+1)$-st mode after obtaining a multiplicative error of $(1+\eta)$ factor in the i-th mode. Imagine that we flatten the first $(i+1)$-modes as a $d \times d^i$ matrix. It is tempting to view one's estimate in the $(i+1)$-st mode as providing an approximation to the 1-norm of the *vector of estimates* of rows produced by $\mathbf{S}^i$. Since we hash the rows (the first i modes) into buckets as in a CountSketch structure, a heavy row in a bucket is perturbed by some small noise and we need to claim that this small perturbation only incurs a small error in the estimate of the row by $\mathbf{S}^i$. An issue arises that a small perturbation in 1-norm on the first i modes may appear larger for a heavy row on the first $(i-1)$ modes, or, equivalently, the first $(i+1)$ modes can tolerate a constant-factor smaller perturbation under $\mathbf{S}^{i+1}$ than the first i modes under $\mathbf{S}^i$, and thus $\mathbf{S}^{i+1}$ needs to use a constant-factor more number of buckets than $\mathbf{S}^i$ to reduce the error in each bucket, resulting in the same $2^{O(q^2)}$ factor in the overall sketching dimension. To see that the shrinking perturbation on higher modes is indeed possible, see Figures 2 and 3 for example. In Figure 2, the $d \times d$ matrix has norm $\Theta(d)$ and exactly one ε-heavy row. To recover the heavy row, the rows are hashed into $1/\varepsilon^2$ buckets and the heavy row is combined with exactly one value of $\varepsilon^2 d$ at the specified entry in

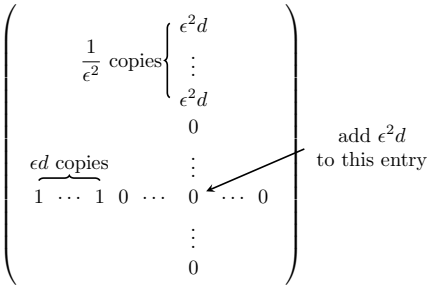

Figure 2: Hard instance for the attempted improvement when $q = 2$. The algorithm first hashes rows into buckets.

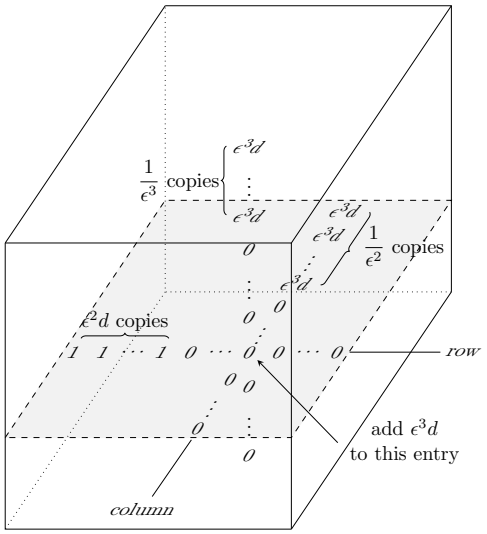

Figure 3: Hard instance for the attempted improvement when $q = 3$. The algorithm first hashes horizontal slices into buckets (parallel to the shaded slice), then the sub-algorithm for each bucket (which contains a linear combination of horizontal slices) hashes rows into buckets.

some bucket. Note that the entry is an ε-heavy hitter in the combined row. Adding a value of $\varepsilon^2 d$ to the specified entry is only an ε^2-factor perturbation to the overall matrix but an ε-factor perturbation to the bucket and a constant-factor perturbation to that entry. Similarly, in Figure 3, adding a value of $\varepsilon^3 d$ to the specified entry is only an ε^3-factor perturbation to the overall $d \times d \times d$ cube but an ε^2-factor perturbation to the only ε-heavy slice (shaded) and an ε-factor perturbation to the only ε-heavy row on that slice.

It is important to note that the work of Braverman and Ostrovsky Braverman and Ostrovsky (2010a) also applies ℓ_1-sketches in the context of tensor products. However, the subroutines used in Braverman and Ostrovsky (2010a) define both level sets and subsampling rates in power of $1 + \varepsilon$, and ε can be shown to become polynomially smaller in each recursive step, and consequently, when iterating this process for a general tensor of order q, at the base level it requires a $(1 + \varepsilon^{2^q})$-approximation to the relevant quantities, resulting in a doubly exponential $\Omega(1/\varepsilon^{2^q})$ amount of memory. Removing the $1/\varepsilon^{2^q}$ term from their space complexity does not appear to be straightforward Braverman (2020). In contrast, our algorithm is a more direct analogue of TENSORSKETCH Pagh (2013); Pham and Pagh (2013); Avron et al. (2014); Ahle et al. (2020) but for the 1-norm, and admits a simpler analysis, leading to a singly exponential sketching dimension as well as a singly exponential memory bound in a data stream.

Given the simplicity and modular components of our algorithm, we can extend it to any distance measure with a (1) small so-called *Rademacher dimension*, a (2) black box sketching algorithm, and (3) an approximate triangle inequality.

1.2.4. POLYNOMIAL-SIZED SUBSPACE EMBEDDINGS

In order to obtain even better oblivious subspace embeddings into ℓ_1, we consider the case when the input matrix $\mathbf{A}$ itself has i.i.d. entries. This models settings in statistics with random design matrices for regression, and our results can be viewed from the lens of average-case complexity. The important property from the distribution on each entry of $\mathbf{A}$ is its tail.

We give the intuition for our improved upper bounds when $\mathbf{A}$ is a matrix of i.i.d. Cauchy random variables. We obtain an $O((\log n)/\log d)$-approximation by simply using a CountSketch matrix $\mathbf{S}$ with $\mathrm{poly}(d)$ rows. When n is at most a polynomial in d, this gives an $O(1)$-approximation, bypassing the $\Omega(d/\log^2 d)$ lower bound of Wang and Woodruff (2019) for arbitrary input matrices $\mathbf{A}$. The idea is that by looking at the rows of $\mathbf{A}$ containing the largest $\mathrm{poly}(d)$ entries in $\mathbf{A}$ - call this submatrix of rows $\mathbf{A}_{top}$ - then we can show $\|\mathbf{A}_{top}\mathbf{x}\|_1 \geqslant n(\log d)\|\mathbf{x}\|_1$ for all $\mathbf{x}$. On the other hand, one can show that for any x, $\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{A}_{top}\mathbf{x}\|_1 + (n\log n)\|\mathbf{x}\|_1$, by concentration bounds applied to the rows not containing a large entry. Finally, we use that (1) CountSketch does not increase the 1-norm of any vector it is applied to, and (2) it perfectly hashes the rows in $\mathbf{A}_{top}$. Putting these statements together gives us an $O((\log n)/\log d)$-approximation.

We also give a number of lower bounds, showing that our algorithms for random $\mathbf{A}$ are also nearly optimal in their sketching dimension. These results are presented in Section E.

1.3. Additional Related Work

Our focus is on linear oblivious maps. Besides being a fundamental mathematical object, such maps are essential for the data stream and distributed models above, allowing for very fast update time under updates. There are other, non-oblivious embeddings for n points in ℓ_1, achieving $O(n/\varepsilon^2)$ dimensions Newman and Rabinovich (2010); Schechtman (1987); Talagrand (1990), which is nearly optimal Charikar and Sahai (2002); Brinkman and Charikar (2005); Andoni et al. (2011). See also Cohen and Peng (2015); Talagrand (1990) for non-oblivious subspace embeddings based on Lewis weights.

For oblivious subspace embeddings, one can achieve $O(d\log d)$ distortion with a sketching dimension of $O(d\log d)$ using a matrix of Cauchy random variables Sohler and Woodruff (2011). This is a significantly larger distortion than the distortion we seek here. It does not contradict the lower bound of Wang and Woodruff (2019) which grows roughly as $\Omega(d/\log^2 r)$, where r is the sketching dimension.

Acknowledgments

We thank anonymous reviewers for their feedback, and T. Yasuda thanks Manuel Fernandez for useful discussions. D. Woodruff and T. Yasuda thank partial support from a Simons Investigator Award. Y. Li was supported in part by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 141–160. SIAM, 2020.

A. Andoni, K. D. Ba, P. Indyk, and D. Woodruff. Efficient sketches for earth-mover distance, with applications. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 324–330, 2009.

Alexandr Andoni, Moses Charikar, Ofer Neiman, and Huy L. Nguyen. Near linear lower bound for dimension reduction in L1. In Rafail Ostrovsky, editor, *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 315–323. IEEE Computer Society, 2011. doi: 10.1109/FOCS.2011.87. URL https://doi.org/10.1109/FOCS.2011.87.

Haim Avron, Huy L. Nguyen, and David P. Woodruff. Subspace embeddings for the polynomial kernel. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2258–2266, 2014. URL http://papers.nips.cc/paper/5240-subspace-embeddings-for-the-polynomial-kernel.

Guus Balkema and Paul Embrechts. Linear regression for heavy tails. *Risks*, 6(3):93, 2018.

Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for ℓ_p-low rank approximation. In *SODA*, pages 747–766. SIAM, 2019.

Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. Simpler algorithm for estimating frequency moments of data streams. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pages 708–713. ACM Press, 2006. URL http://dl.acm.org/citation.cfm?id=1109557.1109634.

Jean Bourgain, Joram Lindenstrauss, V Milman, et al. Approximation of zonoids by zonotopes. *Acta mathematica*, 162:73–141, 1989.

Christos Boutsidis, David P Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 236–249, 2016.

Vladimir Braverman. personal communication, 2020.

Vladimir Braverman and Rafail Ostrovsky. Measuring independence of datasets. arXiv:0903.0034 [cs.DS]. This is the full version of the conference version that appears in STOC'10, 2009. URL https://arxiv.org/abs/0903.0034.

Vladimir Braverman and Rafail Ostrovsky. Measuring independence of datasets. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 271–280, 2010a.

Vladimir Braverman and Rafail Ostrovsky. Zero-one frequency laws. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 281–290, 2010b.

Vladimir Braverman, Kai-Min Chung, Zhenming Liu, Michael Mitzenmacher, and Rafail Ostrovsky. AMS without 4-wise independence on product domains. In *27th International Symposium on Theoretical Aspects of Computer Science, STACS 2010, March 4-6, 2010, Nancy, France*, pages 119–130, 2010.

Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in l1. *Journal of the ACM (JACM)*, 52(5):766–788, 2005.

Moses Charikar and Amit Sahai. Dimension reduction in the/spl lscr//sub 1/norm. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 551–560. IEEE, 2002.

Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *ICALP*, volume 2380 of *Lecture Notes in Computer Science*, pages 693–703. Springer, 2002.

Kenneth L Clarkson. Subgradient and sampling algorithms for l 1 regression. In *Symposium on Discrete Algorithms: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, volume 23, pages 257–266, 2005.

Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214, 2009.

Kenneth L. Clarkson and David P. Woodruff. Sketching for *M*-estimators: A unified approach to robust regression. In *SODA*, pages 921–939. SIAM, 2015.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. In *SODA*, pages 466–477. SIAM, 2013.

Michael B Cohen and Richard Peng. Lp row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 183–192, 2015.

Graham Cormode, Minos Garofalakis, Peter J Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1–3): 1–294, 2012.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson: Lindenstrauss transform. In *STOC*, pages 341–350. ACM, 2010.

Y. Dodge. *L1-statistical Analysis and Related Methods*. North Holland, 1992. ISBN 9780444894441. URL https://books.google.com/books?id=mBzvAAAAMAAJ.

Joan Feigenbaum, Sampath Kannan, Martin J Strauss, and Mahesh Viswanathan. An approximate l 1-difference algorithm for massive data streams. *SIAM Journal on Computing*, 32(1):131–151, 2002.

Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 630–649, 2010.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006a. doi: 10.1145/1147954.1147955. URL https://doi.org/10.1145/1147954.1147955.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006b.

Piotr Indyk and Andrew McGregor. Declaring independence via the sketching of sketches. In *SODA*, volume 8, pages 737–745, 2008.

Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In Harold N. Gabow and Ronald Fagin, editors, *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 202–208. ACM, 2005. doi: 10.1145/1060590.1060621. URL https://doi.org/10.1145/1060590.1060621.

Piotr Indyk, Nick Koudas, and Shanmugavelayutham Muthukrishnan. Identifying representative trends in massive time series data sets using sketches. In *26th International Conference on Very Large Data Bases, VLDB 2000*, pages 363–372, 2000.

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

Khaled Labib and V Rao Vemuri. A hardware-based clustering approach for anomaly detection. *International Journal of Network Security*, 2005.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 633–638. IEEE Computer Society, 2017. doi: 10.1109/FOCS.2017.64. URL https://doi.org/10.1109/FOCS.2017.64.

Kenneth D Lawrence. *Robust regression: analysis and applications*. Routledge, 2019.

James R Lee. Lecture 7-8 martingales and Azuma's inequality, 2016. URL https://homes.cs.washington.edu/~jrl/teaching/cse525au16/lectures/lecture7.pdf.

Roie Levin, Anish Prasad Sevekari, and David P. Woodruff. Robust subspace approximation in a stream. In *NeurIPS*, pages 10706–10716, 2018.

Rake& Agrawal King-lp Lin and Harpreet S Sawhney Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *Proceeding of the 21th International Conference on Very Large Data Bases*, pages 490–501. Citeseer, 1995.

Arvind V. Mahankali and David P. Woodruff. Optimal ℓ_1 column subset selection and a fast PTAS for low rank approximation. In *SODA*, pages 560–578. SIAM, 2021.

Andrew McGregor and Hoa T. Vu. Evaluating bayesian networks via data streams. In Dachuan Xu, Donglei Du, and Dingzhu Du, editors, *Computing and Combinatorics*, pages 731–743, Cham, 2015. Springer International Publishing.

Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100, 2013.

Shanmugavelayutham Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.

Jelani Nelson and Huy L Nguyen. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*, pages 883–894. Springer, 2014.

Ilan Newman and Yuri Rabinovich. On cut dimension of ℓ_1 metrics and volumes, and related sparsification techniques. *CoRR*, abs/1002.3541, 2010.

John P. Nolan. *Univariate Stable Distributions*. Springer, Cham, 2018.

Rasmus Pagh. Compressed matrix multiplication. *ACM Trans. Comput. Theory*, 5(3):9:1–9:17, 2013. doi: 10.1145/2493252.2493254. URL https://doi.org/10.1145/2493252.2493254.

Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy, editors, *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 239–247. ACM, 2013. doi: 10.1145/2487575.2487591. URL https://doi.org/10.1145/2487575.2487591.

Alexander R Pruss. Comparisons between tail probabilities of sums of independent symmetric random variables. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 33, pages 651–671. Elsevier, 1997.

Gideon Schechtman. More on embedding subspaces of l_p in l_r^n. *Compositio Mathematica*, 61(2): 159–169, 1987.

Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *STOC*, pages 755–764. ACM, 2011.

Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701, 2017.

Michel Talagrand. Embedding subspaces of l1 into ln 1. *Proceedings of the American Mathematical Society*, pages 363–369, 1990.

Elad Verbin and Qin Zhang. Rademacher-sketch: A dimensionality-reducing embedding for sum-product norms, with an application to earth-mover distance. In *International Colloquium on Automata, Languages, and Programming*, pages 834–845. Springer, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ruosong Wang and David P. Woodruff. Tight bounds for ℓ_p oblivious subspace embeddings. In *SODA*, pages 1825–1843. SIAM, 2019.

David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.

David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 546–567. JMLR.org, 2013. URL http://proceedings.mlr.press/v30/Woodruff13.html.

Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *NeurIPS*, pages 1084–1094, 2018.

Appendix A. Preliminaries

A.1. Subspace embeddings

We record some results in the literature that are standard ingredients in the construction and analysis of subspace embeddings. We first recall the CountSketch construction.

Definition A.1 (CountSketch Charikar et al. (2002)) *CountSketch is a distribution over $r \times n$ matrices that samples a random matrix $\mathbf{S}$ as follows.*

- *Let $H : [n] \to [r]$ be a random hash function, so that $H(i) = r'$ for $r' \in [r]$ with probability $1/r$.*
- *For each $i \in [n]$, let $\Lambda_i \sim \{\pm 1\}$.*
- *$\mathbf{S}$ is an $r \times n$ matrix taking values in $\{-1, 0, 1\}$ such that $\mathbf{S}_{H(i),i} = \Lambda_i$ for each $i \in [n]$ and 0s everywhere else.*

Remark A.2 *The CountSketch construction originated in the data stream literature Charikar et al. (2002) and has been successfully applied to problems in numerical linear algebra in works such as Dasgupta et al. (2010); Clarkson and Woodruff (2017, 2015).*

The next lemma is useful for net arguments:

Lemma A.3 (Net argument) *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathcal{S} \coloneqq \{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{Ax}\| = 1\}$. Let $\varepsilon \in (0, 1/2)$.*

- *There exists an ℓ_1 ε-net $\mathcal{N}$ of size at most $(3/\varepsilon)^d = \exp(d\log(3/\varepsilon))$ over $\mathcal{S}$, that is, for every $\mathbf{y} \in \mathcal{S}$ there exists a $\mathbf{y}' \in \mathcal{N}$ such that $\|\mathbf{y} - \mathbf{y}'\|_1 \leqslant \varepsilon$ Bourgain et al. (1989).*
- *Let $\mathbf{y} \in \mathcal{S}$. Then, $\mathbf{y} = \sum_{i=0}^{\infty} \mathbf{y}^{(i)}$ where each nonzero $\mathbf{y}^{(i)}$ has $\mathbf{y}^{(i)}/\|\mathbf{y}^{(i)}\|_1 \in \mathcal{N}$ and $\|\mathbf{y}^{(i)}\|_1 \leqslant \varepsilon^i$ (Wang and Woodruff, 2019, implicit in Theorem 3.5).*

The next lemma uses a standard balls and bins martingale argument (e.g., Lee (2016)) to show concentration for uniquely hashed items. This is used in Clarkson and Woodruff (2015) to analyze the M-sketch.

Lemma A.4 (Concentration for unique hashing) *Let $h : [n] \to [r]$ be a random hash function. Let $S \subseteq T \subseteq [n]$, $p \in (0, 1]$, and $\varepsilon \in (0, 1)$ with $\varepsilon r \geqslant p|T|$. Consider the process that samples each element $i \in [n]$ with probability p and hashes it to a bucket in $[r]$ if it was sampled. Let X be the number of elements $i \in S$ that are sampled and hashed to a bucket containing no other member of T. Then,*

$$\mathbf{Pr}\big(X \geqslant (1-\varepsilon)^2 p|S|\big) \leqslant 2\exp\left(-\frac{\varepsilon^2}{12}p|S|\right).$$

Proof The proof is deferred to Appendix F. ■

Theorem A.5 (Improvement of Theorem 3.5, Wang and Woodruff (2019)) *Let $\varepsilon \in (0, 1)$, $r = \exp(\exp(O(d\varepsilon^{-2}\log\varepsilon^{-1} + \varepsilon^{-2}\log\delta^{-1})))$, and let $\mathbf{S}$ be an $r \times n$ matrix of i.i.d. Cauchys. Then for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,*

$$\mathbf{Pr}\{(1-\varepsilon)\|\mathbf{Ax}\|_1 \leqslant \|\mathbf{SAx}\|_1 \leqslant (1+\varepsilon)\|\mathbf{Ax}\|_1\} \geqslant 1 - \delta.$$

Proof The proof is deferred to Appendix F. ■

Remark A.6 *Note that the above dense sketch preserves an arbitrary fixed vector with probability at least $1 - \delta$ using a sketching dimension of $2^{1/\delta}$. Thus, for preserving the 1-norm of n arbitrary vectors, it suffices to set $\delta = O(1/n)$. On the other hand, the lower bound argument of (Wang and Woodruff, 2019, Theorem 1.1) proves a distortion lower bound for sketching matrices that preserve even just the columns of the input matrix $\mathbf{A}$. Thus, we can place our n vectors along the columns of a matrix, so that for constant distortion, a sketch needs r dimensions, for*

$$\frac{n}{\log^2 r} = O(1) \implies r = \Omega(2^{\sqrt{n}}).$$

Appendix B. Singly Exponential $(1+\varepsilon)$ ℓ_1 Subspace Embeddings

In this section, we prove the following theorem:

Theorem B.1 *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Then there exists a sparse oblivious ℓ_1 subspace embedding $\mathbf{S}$ into r dimensions with*

$$r = \mathrm{poly}(d, \varepsilon^{-1}, \delta^{-1}, \log n)^{d/\delta\varepsilon}$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\Pr_{\mathbf{S}}\Big\{\forall \mathbf{x} \in \mathbb{R}^d, (1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_1, \Big\} \geqslant 1-\delta.$$

Our main contribution towards proving this result is in showing the "no dilation" direction $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_1$. The "no contraction" direction of $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \geqslant (1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_1$ direction was already known in Clarkson and Woodruff (2015), and we defer the details of handling our minor changes to Appendix G.

If we settle for dense embeddings, then we are able to get an improved sketching dimension that is independent of n by first applying the dense ℓ_1 subspace embedding of Theorem A.5, which maps our subspace down to a subspace of dimension independent of n and preserves 1-norms up to a $(1+\varepsilon)$ factor distortion:

Corollary B.2 *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Then there exists an oblivious ℓ_1 subspace embedding $\mathbf{S}$ into r dimensions with*

$$r = \exp\Big(\tilde{O}(d^2/\delta\varepsilon^3)\Big)$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\Pr_{\mathbf{S}}\Big\{\forall \mathbf{x} \in \mathbb{R}^d, (1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_1\Big\} \geqslant 1-\delta.$$

Proof By applying the sketch of Theorem A.5 first, we can take $\log\log n \leqslant d/\delta\varepsilon^2$. Then, the bounds for Theorem B.1 yield the desired result. ■

By a known lower bound in Theorem 1.1 of Wang and Woodruff (2019), the dependence on d is optimal up to polynomial factors in the exponent.

B.1. The embedding

We first collect constants that will be used. The constants can all be written in terms of the dimensions n and d of the input matrix, the accuracy parameter ε, and the failure rate δ.

Definition B.3 (Useful constants)

$$
\begin{aligned}
h_{\max} &\coloneqq \log_2(n/\varepsilon) && = O(\log(n/\varepsilon)) && \textit{Sampling levels}\\
q_{\max} &\coloneqq \log_2(ndh_{\max}/\delta\varepsilon) && = O(\log(nd/\varepsilon)) && \textit{Weight classes}\\
\alpha &\coloneqq 2\exp(d\log(3/\varepsilon))q_{\max}/\delta && = O\left(\frac{\exp(d/\varepsilon)\log(nd/\varepsilon)}{\delta}\right) && \textit{Net union bounding}\\
m_{\text{crowd}} &\coloneqq 300\frac{d^{11}}{\varepsilon^9\delta^4}\log^5(n) && = O(\operatorname{poly}(d,\varepsilon^{-1},\delta^{-1},\log n)) && \textit{Overcrowding hash buckets}\\
B &\coloneqq (m_{\text{crowd}}h_{\max}q_{\max}/\delta)^{d/\delta\varepsilon} && = O\Big(\operatorname{poly}(d,\varepsilon^{-1},\delta^{-1},\log n)^{d/\delta\varepsilon}\Big) && \textit{Branching factor}\\
N_0 &\coloneqq \frac{12B^u q_{\max}}{\varepsilon^3}\log\alpha && && \textit{Hash buckets in 0th level}\\
N &\coloneqq B\frac{8d^2\log d}{\varepsilon^6}q_{\max}(\log\alpha)\left(\log\frac{B}{\varepsilon}\right) && = O(B\log n\operatorname{poly}(d,\varepsilon^{-1})) && \textit{Hash buckets per level}
\end{aligned}
$$

As described in the introduction, the construction of our embedding is essentially a variant of M-sketch Clarkson and Woodruff (2015). However, instead of using fixed subsampling rates of $1/\operatorname{poly}(d)$, we use randomized subsampling rates which drop off geometrically by factors of $B = O\big(\operatorname{poly}(d,\varepsilon^{-1},\delta^{-1},\log n)^{d/\delta\varepsilon}\big)$.

Definition B.4 *Let $u \sim [0,1]$ and define subsampling rates*

$$p_h \coloneqq B^{-(u+h-1)}$$

for each $h \in [h_{\max}]$.

Definition B.5 *For each $i \in [n]$ and $h \in [h_{\max}]$, let*

$$b_{i,h} \coloneqq \begin{cases} 1 & \textit{w.p. } p_h \\ 0 & \textit{w.p. } 1-p_h \end{cases},$$

and let $m_h \coloneqq \sum_{i\in[n]} b_{i,h}$.

Definition B.6 *For each $i \in [n]$, let $\Lambda_i \sim \{\pm 1\}$. Let $H_0 : [n] \to [N_0]$ and $H_h : [m_h] \to [N]$ for each $h \in [h_{\max}]$ be a random hash functions.*

Definition B.7 (Random-boundary M-sketch) *Let $\mathbf{C}^{(0)}$ be an $N_0 \times n$ CountSketch matrix (Definition A.1) with random signs Λ_i and hash function H_0, that is,*

$$\mathbf{C}^{(0)}_{H_0(i),i} \coloneqq \Lambda_i$$

for every $i \in [n]$ and 0s everywhere else. For each $h \in [h_{\max}]$, let $\mathbf{S}^{(h)}$ be the $m_h \times n$ scaled sampling matrix given by

$$\mathbf{e}_j^\top \mathbf{S}^{(h)} \mathbf{e}_i = \begin{cases} \frac{1}{p_h} & j = \sum_{i'\in[i]} b_{i',h} \\ 0 & \textit{otherwise} \end{cases}.$$

For each $h \in [h_{\max}]$, let $\mathbf{C}^{(h)}$ be an $N \times m_h$ CountSketch matrix with random signs Λ_i and hash function H_h, that is,

$$\mathbf{C}^{(h)}_{H_h(i),i} \coloneqq \Lambda_i$$

for each $\mathbf{y}_i$ that was sampled, i.e., $b_{i,h} = 1$, and 0s everywhere else. Then, our random-boundary M-sketch is given by

$$\mathbf{S} \coloneqq \begin{pmatrix} \mathbf{C}^{(0)} \\ \mathbf{C}^{(1)}\mathbf{S}^{(1)} \\ \mathbf{C}^{(2)}\mathbf{S}^{(2)} \\ \vdots \\ \mathbf{C}^{(h_{\max})}\mathbf{S}^{(h_{\max})} \end{pmatrix}.$$

B.2. Notation for analysis

We first recall some notation from the analysis of M-sketch in Clarkson and Woodruff (2015), as well as a few other definitions.

Definition B.8 *Let $\mathbf{y} \in \mathbb{R}^n$ be a unit ℓ_1 vector and let $q \in \mathbb{N}$. We define weight classes*

$$W_q(\mathbf{y}) \coloneqq \left\{\mathbf{y}_i : 2^{-q} \leqslant |\mathbf{y}_i| \leqslant 2^{1-q}\right\}.$$

When the $\mathbf{y}$ is clear from context, we simply write W_q for brevity. For a set $Q \subseteq \mathbb{N}$, we write

$$W_Q \coloneqq \bigcup_{q \in Q} W_q.$$

We also write $|W_q|$ for the size of W_q and

$$\|W_q\|_1 \coloneqq \sum_{y \in W_q} |y|.$$

Definition B.9 *For $h \in [h_{\max}]$ and $k \in [N]$, we write $L_{h,k}$ for the multiset of elements that get sampled and hashed to the kth bucket in the hth level.*

We briefly digress to recall ℓ_1 leverage score vectors.

Definition B.10 (ℓ_1 well-conditioned basis (Definition 2, Clarkson et al. (2013), see also Dasgupta et al. (2009))) *A basis $\mathbf{U}$ for the range of an $n \times d$ matrix $\mathbf{A}$ is (α, β)-conditioned if $\|U\|_1 \leqslant \alpha$ and for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_\infty \leqslant \beta\|\mathbf{U}\mathbf{x}\|_1$. We say that $\mathbf{U}$ is well-conditioned if α and β are low-degree polynomials d, independent of n. It is known that an Auerbach basis for $\mathbf{A}$ is $(d, 1)$-conditioned.*

Definition B.11 (ℓ_1 leverage scores (Definition 3, Clarkson et al. (2013))) *Given a $(d, 1)$-conditioned basis $\mathbf{U}$ (see Definition B.10) for the column space of $\mathbf{A} \in \mathbb{R}^{n \times d}$, define the vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ of normalized ℓ_1 leverage scores of $\mathbf{A}$ to be*

$$\boldsymbol{\lambda}_i \coloneqq \frac{\|\mathbf{e}_i^\top \mathbf{U}\|_1}{d}.$$

Remark B.12 *As noted in Clarkson et al. (2013), the ℓ_1 leverage scores are not defined uniquely. We also note that for convenience of notation, our normalization of the leverage scores is off by a factor of d from standard definitions in the literature.*

In our analysis, we consider weight classes $W_q(\boldsymbol{\lambda})$ of the ℓ_1 leverage score vector $\boldsymbol{\lambda}$. For each weight class W_q, we set

$$h_q \coloneqq \lfloor \log_B |W_q| \rfloor$$

so that $B^{h_q} \leqslant |W_q| < B^{h_q+1}$.

Definition B.13 *For a pair $(h, q) \in [h_{\max}] \times \mathbb{N}$ and an interval I, define the event*

$$\mathcal{E}_{h,q}(I) \coloneqq \{p_h |W_q(\boldsymbol{\lambda})| \in I\}$$

in which sampling the weight class $W_q(\boldsymbol{\lambda})$ at rate p_h has an expected number of items in the window I.

Definition B.14 (Scaled leverage score samples) *For each $(h, q) \in [h_{\max}] \times [q_{\max}]$ and an interval I, define the random variables*

$$\mathcal{S}_{h,q} \coloneqq \frac{1}{p_h} \sum_{\boldsymbol{\lambda}_i \in W_q} b_{i,h} \boldsymbol{\lambda}_i$$

$$\mathcal{T}_{h,q}(I) \coloneqq \frac{1}{p_h} \sum_{\boldsymbol{\lambda}_i \in W_q} b_{i,h} \boldsymbol{\lambda}_i \mathbb{1}(\mathcal{E}_{h,q}(I))$$

In the following sections, we give upper bounds on the mass of the sketch depending on the weight class of the leverage scores that we look at. We have the following intervals:

- **Dead levels** $p_h |W_q(\boldsymbol{\lambda})| \in [0, \delta / h_{\max} q_{\max})$: In this interval, we sample none of these entries with high probability.

- **Badly concentrated levels** $p_h |W_q(\boldsymbol{\lambda})| \in [\delta / h_{\max} q_{\max}, m_{\text{crowd}})$: The expected mass of leverage scores coming from this level is at most $O(\varepsilon/d)$, which means that with constant probability, the mass contribution for all subspace vectors is $O(\varepsilon)$.

- **Golidlocks levels** $p_h |W_q(\boldsymbol{\lambda})| \in [m_{\text{crowd}}, B m_{\text{crowd}})$: In this interval, we can show that the mass contribution is at most a $(1 + \varepsilon)$ factor more than the expected mass coming from this interval with high probability. This level is counted only once, since the size of the interval is less than a B factor.

- **Oversampled levels** $p_h |W_q(\boldsymbol{\lambda})| \in [B m_{\text{crowd}}, \infty)$: In this interval, we sample so many of these entries that it overcrowds the CountSketch hash buckets, which makes the mass contribution at most an ε fraction due to the random sign cancellations.

B.3. Bounding badly concentrated levels

For levels with expected mass in the interval $[1/\alpha, \log \alpha]$ at subsampling rate p_h, we cannot hope to reason about the mass contribution of this level with high enough probability to union bound over a net, since we need expectation at most $1/\alpha$ for the level to get completely missed by the sampling, and we need at least $\log \alpha$ in order to get concentration. However, we show that because of our randomization of subsampling rates, the leverage score mass contribution from these rows is only an $O(\varepsilon/d)$ fraction of the total mass of the leverage scores in expectation, which means it is only an $O(\varepsilon)$ fraction of the total mass of any subspace vector with constant probability by a combination of properties of leverage scores and a Markov bound.

Lemma B.15 (Randomized sampling rates) *Let $\delta' \in (0,1)$, let $0 < a < 1$ and $b > 1$, and let $B' \coloneqq (b/a)^{1/\delta'}$. Let $u \sim [0,1]$, $p = B'^{-u}$, and let $t \in \mathbb{R}$. Then,*

$$\mathbf{Pr}(pt \in [a,b]) \leqslant \begin{cases} 0 & \text{if } t \geqslant b \text{ or } B't \leqslant a \\ \delta & \text{otherwise.} \end{cases}$$

Proof The first bound follows from the fact that $t = B'^0 t \leqslant pt \leqslant B'^1 t = B't$. For the second bound, we calculate

$$\mathbf{Pr}(pt \in [a,b]) = \mathbf{Pr}(u \in \log_{B'} t + [-\log_{B'} b, -\log_{B'} a]) \leqslant \log_{B'}(b/a) = \delta' \frac{\log(b/a)}{\log(b/a)} = \delta'.$$

■

Corollary B.16 *For every $h \in [h_{\max}]$ and $q \in [q_{\max}]$,*

$$\mathbf{Pr}(\mathcal{E}_i([\delta/h_{\max}q_{\max}, m_{\text{crowd}}))) = \mathop{\mathbf{Pr}}_u(p_h|W_q| \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})) \leqslant \begin{cases} 0 & \text{if } h \notin \{h_q, h_q+1\} \\ \frac{\delta\varepsilon}{d} & \text{otherwise} \end{cases}.$$

Proof Note that for $h \geqslant h_q + 2$,

$$B^{-h}|W_q| \leqslant B^{-h+h_q+1} \leqslant B^{-1} \leqslant \frac{\delta}{h_{\max}q_{\max}}$$

and for $h \leqslant h_q - 1$,

$$B^{-h}|W_q| \geqslant B^{-h+h_q} \geqslant B^1 \geqslant m_{\text{crowd}}$$

so for $h \notin \{h_q, h_q+1\}$,

$$\mathop{\mathbf{Pr}}_u(p_h|W_q| \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})) = \mathop{\mathbf{Pr}}_u\Big(B^{-u}\Big(B^{-h}|W_q|\Big) \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})\Big) = 0.$$

On the other hand, for $h \in \{h_q, h_q+1\}$,

$$\mathop{\mathbf{Pr}}_u(p_h|W_q| \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})) \leqslant \frac{\delta\varepsilon}{d}$$

by Lemma B.15. ■

Note that by Corollary B.16, $\mathcal{E}_{h,q}([\delta/h_{\max}q_{\max}, m_{\text{crowd}}))$ has nonzero probability for only $h \in \{h_q, h_q+1\}$.

Lemma B.17 (Expected mass of bad leverage scores)

$$\mathop{\mathbf{E}}_{u,b}\left(\sum_{q\in[q_{\max}]}\sum_{h\in[h_{\max}]}\mathcal{T}_{h,q}([\delta/h_{\max}q_{\max}, m_{\text{crowd}}))\right) \leqslant \frac{4\delta\varepsilon}{d}.$$

Proof Let $I \coloneqq [\delta/h_{\max}q_{\max}, m_{\text{crowd}})$. Then,

$$\begin{aligned}
\mathop{\mathbf{E}}_{u,b}\left(\sum_{q\in[q_{\max}]}\sum_{h\in[h_{\max}]}\mathcal{T}_{h,q}(I)\right) &= \mathop{\mathbf{E}}_{u,b}\left(\sum_{q\in[q_{\max}]}\sum_{h\in\{h_q,h_q+1\}}\mathcal{T}_{h,q}(I)\right)\\
&= \sum_{q\in[q_{\max}]}\sum_{h\in\{h_q,h_q+1\}}\mathop{\mathbf{E}}_{u,b}\mathcal{T}_{h,q}(I)\\
&= \sum_{q\in[q_{\max}]}\sum_{h\in\{h_q,h_q+1\}}\sum_{\boldsymbol{\lambda}_i\in W_q}\mathop{\mathbf{E}}_{u,b}\left[\frac{1}{p_h}b_{i,h}\boldsymbol{\lambda}_i\mathbb{1}(\mathcal{E}_{h,q}(I))\right]\\
&\leqslant \sum_{q\in[q_{\max}]}\sum_{h\in\{h_q,h_q+1\}}\sum_{\boldsymbol{\lambda}_i\in W_q}2^{1-q}\mathop{\mathbf{E}}_{u}(\mathbb{1}(\mathcal{E}_{h,q}(I)))\\
&\leqslant \sum_{q\in[q_{\max}]}\sum_{h\in\{h_q,h_q+1\}}\sum_{\boldsymbol{\lambda}_i\in W_q}2^{1-q}\frac{\delta\varepsilon}{d}\\
&= \sum_{q\in[q_{\max}]}2^{2-q}|W_q|\frac{\delta\varepsilon}{d}\\
&\leqslant \frac{4\delta\varepsilon}{d}.
\end{aligned}$$

■

Lemma B.18 *For any $\mathbf{x}\in\mathbb{R}^d$ and $i\in[n]$, we have that*

$$\frac{|\mathbf{e}_i^\top\mathbf{A}\mathbf{x}|}{\|\mathbf{A}\mathbf{x}\|_1} \leqslant d\boldsymbol{\lambda}_i$$

Proof Let $\mathbf{y}\in\mathbb{R}^d$ be such that $\mathbf{A}\mathbf{x}=\mathbf{U}\mathbf{y}$. Then,

$$\frac{|\mathbf{e}_i^\top\mathbf{A}\mathbf{x}|}{\|\mathbf{A}\mathbf{x}\|_1} = \frac{|\mathbf{e}_i^\top\mathbf{U}\mathbf{y}|}{\|\mathbf{U}\mathbf{y}\|_1} \leqslant \frac{\|\mathbf{e}_i^\top\mathbf{U}\|_1\|\mathbf{y}\|_\infty}{\|\mathbf{y}\|_\infty} = \|\mathbf{e}_i^\top\mathbf{U}\|_1 = d\boldsymbol{\lambda}_i$$

where the first inequality follows from properties of well-conditioned bases. ■

B.4. Bounding Goldilocks levels

In this level, the expected sampled mass is large enough to get concentration, but not large enough to overflow the hash buckets of the CountSketch. In this level, we show that the mass contribution is at most a $(1+\varepsilon)$ factor more than the expected mass. The main idea for getting concentration here is using the bounds on the leverage scores to bound outliers, and using a Bernstein bound to get concentration on the rest of the entries with a good bound on the variance.

Definition B.19 *Define $\mathbf{A}^{(q)}$ to be the $n \times d$ matrix formed by taking the rows of $\mathbf{A}$ that correspond to leverage scores belonging to weight class $W_q(\boldsymbol{\lambda})$, and 0s everywhere else.*

Lemma B.20 *Let $(h, q) \in [h_{\max}] \times [q_{\max}]$ with $p_h |W_q(\boldsymbol{\lambda})| \geqslant 3d^2\varepsilon^{-4} \log \alpha$ and let $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_1 = 1$. Then with probability at least $1 - 2/\alpha$, we have that*

$$\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 = \sum_{\boldsymbol{\lambda}_i \in W_q(\boldsymbol{\lambda})} \frac{|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}|}{p_h} b_{i,h} \leqslant (1+\varepsilon)\|\mathbf{A}^{(q)}\mathbf{x}\|_1 + 4\varepsilon \|W_q\|_1 \|\mathbf{A}\mathbf{x}\|_1.$$

Proof The average absolute value of an entry of $\mathbf{A}^{(q)}\mathbf{x}$ is $\mu_q \coloneqq \|\mathbf{A}^{(q)}\mathbf{x}\|_1 / |W_q(\boldsymbol{\lambda})|$. Then by averaging, there is at most an ε/d fraction of rows with absolute value greater than $d\mu_q/\varepsilon$. Now for each $\boldsymbol{\lambda}_i \in W_q(\boldsymbol{\lambda})$, define the event

$$\mathcal{F}_i \coloneqq \left\{ |\mathbf{e}_i^\top \mathbf{A}\mathbf{x}| \geqslant \frac{d\mu_q}{\varepsilon} \right\}$$

and the sample

$$X = \sum_{\boldsymbol{\lambda}_i \in W_q} b_{i,h} \mathbb{1}(\mathcal{F}_i).$$

Note that

$$\mathbf{E}\, X = \frac{\varepsilon p_h |W_q|}{d} \geqslant \frac{d \log \alpha}{\varepsilon^4} \geqslant 3 \log \alpha$$

so by the Chernoff bound,

$$\mathbf{Pr}(X \geqslant 2\, \mathbf{E}\, X) \leqslant \exp\left(-\frac{\mathbf{E}\, X}{3}\right) \leqslant \frac{1}{\alpha}.$$

Conditioned on the complement event, the mass contribution from rows i for which $\mathcal{F}_i$ happens is at most

$$\sum_{\boldsymbol{\lambda}_i \in W_q} \frac{|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}|}{p_h} b_{i,h} \mathbb{1}(\mathcal{F}_i) \leqslant 2\frac{\varepsilon p_h |W_q|}{d} \frac{|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}|}{p_h} \leqslant 2\varepsilon \frac{|W_q|}{d} d\boldsymbol{\lambda}_i \|\mathbf{A}\mathbf{x}\|_1 \leqslant 4\varepsilon 2^{-q} |W_q| \|\mathbf{A}\mathbf{x}\|_1 \leqslant 4\varepsilon \|W_q\|_1 \|\mathbf{A}\mathbf{x}\|_1$$

where the second to last inequality follows from Lemma B.18.

We now consider the sample

$$Y = \sum_{\boldsymbol{\lambda}_i \in W_q} Y_i$$

where

$$Y_i \coloneqq \frac{|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}|}{p_h} b_{i,h} \mathbb{1}(\neg \mathcal{F}_i).$$

Note that

$$\begin{aligned}
\mathbf{E}\, Y &\leqslant \|\mathbf{A}^{(q)}\mathbf{x}\|_1 = |W_q| \mu_q \\
Y_i &\leqslant \frac{1}{p_h} \frac{d\mu_q}{\varepsilon} \\
\mathbf{Var}(Y_i) &\leqslant p_h \left(\frac{1}{p_h} \frac{d\mu_q}{\varepsilon} \right)^2 = \frac{1}{p_h} \left(\frac{d\mu_q}{\varepsilon} \right)^2
\end{aligned}$$

Then by Bernstein's inequality,

$$\begin{aligned}\mathbf{Pr}(Y - \mathbf{E}\,Y \geqslant \varepsilon|W_q|\mu_q) &\leqslant \exp\left(-\frac{1}{2}\frac{(\varepsilon|W_q|\mu_q)^2}{|W_q|(d\mu_q/\varepsilon)^2/p_h + (\varepsilon|W_q|\mu_q)(d\mu_q/\varepsilon p_h)/3}\right)\\ &= \exp\left(-\frac{1}{2}\frac{p_h|W_q|\varepsilon^2}{(d/\varepsilon)^2 + d/3}\right) \leqslant \exp\left(-\frac{p_h|W_q|}{3d^2\varepsilon^{-4}}\right) \leqslant \frac{1}{\alpha}.\end{aligned}$$

We conclude by combining the two bounds. ■

B.5. Bounding oversampled levels

When we expect to sample a large enough number of entries per hash bucket from a level, these entries cancel each other out due to the random signs. These levels fall under this criterion.

Lemma B.21 *Let $(h,q) \in [h_{\max}]\times[q_{\max}]$ with $p_h|W_q(\boldsymbol{\lambda})| \geqslant bN$ for $b = 12(\frac{dh_{\max}}{\varepsilon})^2\log(Nh_{\max}q_{\max}/\delta)$. Then with probability at least $1 - 4\delta/h_{\max}q_{\max}$,*

$$\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \leqslant \frac{\varepsilon}{h_{\max}}\|W_q(\boldsymbol{\lambda})\|_1\|\mathbf{A}\mathbf{x}\|_1.$$

Similarly, if $|W_q(\boldsymbol{\lambda})| \geqslant bN_0$, then with probability at least $1 - 4\delta$,

$$\|\mathbf{C}^{(0)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \leqslant \frac{\varepsilon}{h_{\max}}\|W_q(\boldsymbol{\lambda})\|_1\|\mathbf{A}\mathbf{x}\|_1.$$

Proof We just show the first bound since the second is nearly identical. Note that by Lemma B.18, $|\mathbf{e}_i^\top\mathbf{A}^{(q)}\mathbf{x}|/\|\mathbf{A}\mathbf{x}\|_1 \leqslant d\boldsymbol{\lambda}_i \leqslant d2^{1-q}$ for all $\boldsymbol{\lambda}_i \in W_q(\boldsymbol{\lambda})$.

By Chernoff's bound, the probability that a bucket L in level h gets $X = (1 \pm 1/2)p_h|W_q|/N$ elements from W_q is at least

$$\mathbf{Pr}\left(|X - \frac{p_h|W_q|}{N}| \geqslant \frac{1}{2}\frac{p_h|W_q|}{N}\right) \leqslant 2\exp\left(-\frac{(1/2)^2p_h|W_q|}{3}\right) = 2\exp\left(-\frac{p_h|W_q|}{12}\right) \leqslant 2\delta.$$

We condition on this event. Then by Hoeffding's bound, the inner product of m elements $\{a_i\}_{i=1}^m$ in the interval $[d2^{-q}\|\mathbf{A}\mathbf{x}\|_1, d2^{1-q}\|\mathbf{A}\mathbf{x}\|_1]$ with random signs ε_i concentrates around its mean as

$$\begin{aligned}\mathbf{Pr}\left(\sum_{i=1}^m \varepsilon_i a_i > d2^{1-q}\|\mathbf{A}\mathbf{x}\|_1\sqrt{m}\sqrt{\log(Nh_{\max}q_{\max}/\delta)}\right) &\leqslant \exp\left(-\frac{(d2^{1-q}\|\mathbf{A}\mathbf{x}\|_1\sqrt{m}\sqrt{\log(Nh_{\max}q_{\max}/\delta)})^2}{2d^2 2^{2-2q}\|\mathbf{A}\mathbf{x}\|_1^2 m}\right)\\ &\leqslant \frac{\delta}{Nh_{\max}q_{\max}}.\end{aligned}$$

Then by a union bound over N buckets, with probability at least $1 - 2\delta/h_{\max}q_{\max}$, we have for every bucket L at this level that

$$\begin{aligned}\left|\frac{1}{p_h}\sum_{\mathbf{y}_i\in L}\Lambda_i b_{i,h}\mathbf{y}_i\right| &\leqslant d2^{1-q}\|\mathbf{A}\mathbf{x}\|_1\sqrt{X}\sqrt{\log(N/\delta)}\\ &\leqslant \frac{1}{p_h}d2^{1-q}\|\mathbf{A}\mathbf{x}\|_1\sqrt{\frac{3}{2}\frac{p_h|W_q|}{N}}\sqrt{\log(N/\delta)}\\ &\leqslant \frac{1}{p_h}\frac{\varepsilon}{dh_{\max}}d2^{-q}\|\mathbf{A}\mathbf{x}\|_1\frac{p_h|W_q|}{N}\\ &\leqslant \frac{\varepsilon}{h_{\max}}\frac{\|W_q\|_1}{N}\|\mathbf{A}\mathbf{x}\|_1\end{aligned}$$

which gives the desired bound upon summing over the N buckets. The overall success probability is at least $1 - 4\delta/h_{\max}q_{\max}$. ■

B.6. Net argument

In this section, we collect the bounds obtained in previous sections and conclude with a net argument.

Lemma B.22 *With probability at least $1 - 6\delta$, we have for all $\mathbf{x} \in \mathbb{R}^d$ that*

$$\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1\mathbb{1}(\mathcal{F}_{h,q}) \leqslant 5\varepsilon\|\mathbf{A}\mathbf{x}\|_1$$

where

$$\mathcal{F}_{h,q} = \{p_h|W_q| \in [0, m_{\text{crowd}}) \cup [Bm_{\text{crowd}}, \infty)\}$$

Proof We case on $p_h|W_q|$ by intervals $[0, \delta/h_{\max}q_{\max})$, $[\delta/h_{\max}q_{\max}, m_{\text{crowd}})$, and $[Bm_{\text{crowd}}, \infty)$.

- **Dead levels**: First consider the h for which $p_h|W_q| < \delta/h_{\max}q_{\max}$. In this case, the probability that we sample any row corresponding to some $\boldsymbol{\lambda}_i \in W_q$ is at most $p_h|W_q| < \delta/h_{\max}q_{\max}$ by a union bound. Then by a further union bound over all $(h, q) \in [h_{\max}] \times [q_{\max}]$, this category of levels contributes no mass with probability at least $1 - \delta$.
- **Badly concentrated levels**: Consider the subsampling levels with $p_h|W_q| \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})$. By Lemma B.17, the total expected leverage score mass contribution from all such pairs $(h, q) \in [h_{\max}] \times [q_{\max}]$ is at most $4\delta\varepsilon/d$. Then by Markov's inequality, with probability at least $1 - \delta$, the total expected leverage score mass is at most $4\varepsilon/d$. Conditioned on this event, we have that

$$\begin{aligned}&\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\frac{1}{p_h}\sum_{\boldsymbol{\lambda}_i\in W_q} b_{i,h}|\mathbf{e}_i^\top\mathbf{A}\mathbf{x}|\mathbb{1}(\mathcal{E}_{h,q}([\delta/h_{\max}q_{\max}, m_{\text{crowd}}))\\ &\leqslant d\|\mathbf{A}\mathbf{x}\|_1\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\frac{1}{p_h}\sum_{\boldsymbol{\lambda}_i\in W_q} b_{i,h}\boldsymbol{\lambda}_i\mathbb{1}(\mathcal{E}_{h,q}([\delta/h_{\max}q_{\max}, m_{\text{crowd}})) \quad \text{Lemma B.18}\\ &\leqslant d\frac{4\varepsilon}{d}\|\mathbf{A}\mathbf{x}\|_1 = 4\varepsilon\|\mathbf{A}\mathbf{x}\|_1 \quad \text{Lemma B.17}\end{aligned}$$

– **Oversampled levels**: Consider the subsampling levels with $p_h|W_q| \in [Bm_{\text{crowd}}, \infty)$. Note that $Bm_{\text{crowd}} \geqslant bN$ is large enough to apply Lemma B.21. By union bounding and summing over h and q for the result of the lemma, we have that

$$\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]} \|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \mathbb{1}(\mathcal{E}_{h,q}([Bm_{\text{crowd}},\infty)) \leqslant \sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]} \frac{\varepsilon}{h_{\max}}\|W_q\|_1\|\mathbf{A}\mathbf{x}\|_1 \leqslant \varepsilon\|\mathbf{A}\mathbf{x}\|_1$$

with probability at least $1-4\delta$.

We thus conclude by a union bound over the above three events. ■

Lemma B.23 (Tiny weight classes) *Let $q > q_{\max}$. Then with probability at least $1-\delta$, it holds for all $\mathbf{x} \in \mathbb{R}^d$ that*

$$\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}} \|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \leqslant \varepsilon\|\mathbf{A}\mathbf{x}\|_1.$$

Proof For the weight classes $q > q_{\max}$, the total leverage score mass contribution is bounded by

$$\sum_{q>q_{\max}} \|W_q(\boldsymbol{\lambda})\|_1 \leqslant \sum_{q>q_{\max}} 2^{1-q}|W_q| \leqslant \frac{\delta\varepsilon}{dnh_{\max}}\sum_{q>q_{\max}}|W_q| \leqslant \frac{\delta\varepsilon}{dh_{\max}}.$$

Then in expectation, the sum of the scaled leverage score samples (Definition B.14) is bounded by

$$\begin{aligned}
\mathbf{E}\left(\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\mathcal{S}_{h,q}\right) &= \sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\sum_{\boldsymbol{\lambda}_i\in W_q}\mathbf{E}\left(\frac{1}{p_h}b_{i,h}\boldsymbol{\lambda}_i\right)\\
&= \sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|W_q(\boldsymbol{\lambda})\|_1\\
&\leqslant \sum_{h\in[h_{\max}]}\frac{\delta\varepsilon}{dh_{\max}}\\
&= \frac{\delta\varepsilon}{d}.
\end{aligned}$$

Then with probability at least $1-\delta$, the above sum is at most ε/d. We condition on this event. Then, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 &= \sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\sum_{\boldsymbol{\lambda}_i\in W_q}\frac{1}{p_h}b_{i,h}(\mathbf{e}_i^\top\mathbf{A}\mathbf{x})\\
&\leqslant d\|\mathbf{A}\mathbf{x}\|_1\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\sum_{\boldsymbol{\lambda}_i\in W_q}\frac{1}{p_h}b_{i,h}\boldsymbol{\lambda}_i \qquad \text{Lemma B.18}\\
&\leqslant d\|\mathbf{A}\mathbf{x}\|_1\frac{\varepsilon}{d} = \varepsilon\|\mathbf{A}\mathbf{x}\|_1
\end{aligned}$$

as desired. ■

Lemma B.24 *There is an event with probability* $1 - 11\delta$ *such that conditioned on this event, for every* $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{Pr}(\|\mathbf{SAx}\|_1 \leqslant (1+8\varepsilon)\|\mathbf{Ax}\|_1) \geqslant 1 - \frac{2q_{\max}}{\alpha}.$$

Proof By Lemma B.23, the contribution from weight classes $q > q_{\max}$ is at most $\varepsilon\|\mathbf{Ax}\|_1$ with probability at least $1-\delta$. We let this event be $\mathcal{E}_1$ and restrict our attention to $q \leqslant q_{\max}$.

For each $q \in [q_{\max}]$, we bound the mass contribution of rows corresponding to $W_q(\boldsymbol{\lambda})$ at each subsampling level $\{0\} \cup [h_{\max}]$. Note that by Lemma B.22, there is an event $\mathcal{E}_2$ with probability at least $1-6\delta$ such that all levels h, q except for those such that $h = 0$ or $p_h|W_q| \in [m_{\text{crowd}}, Bm_{\text{crowd}})$ are bounded by at most $5\varepsilon\|\mathbf{Ax}\|_1$, so it remains to bound these levels. These are the 0th level of subsampling (i.e., no subsampling) and the Goldilocks levels.

Note that there exists at most one Goldilocks level $h \in [h_{\max}]$ such that $p_h|W_q| \in [m_{\text{crowd}}, Bm_{\text{crowd}})$. In this case, Lemma B.20 applies since $m_{\text{crowd}} \geqslant 3d^2\varepsilon^{-1}\log\alpha$, and we have that

$$\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 = \sum_{\boldsymbol{\lambda}_i \in W_q(\boldsymbol{\lambda})} \frac{|\mathbf{e}_i^\top \mathbf{Ax}|}{p_h} b_{i,h} \leqslant (1+\varepsilon)\|\mathbf{A}^{(q)}\mathbf{x}\|_1 + 4\varepsilon\|W_q\|_1\|\mathbf{Ax}\|_1.$$

with probability at least $1 - 2/\alpha$. If such a Goldilocks subsampling level h exists, then note that

$$p_h|W_q| \geqslant m_{\text{crowd}} \implies |W_q| \geqslant B^{u+h-1}m_{\text{crowd}} \geqslant B^u m_{\text{crowd}} \geqslant bN_0.$$

Then by Lemma B.21, the 0th level of subsampling level contributes mass at most $(\varepsilon/h_{\max})\|W_q(\boldsymbol{\lambda})\|_1$ with probability at least $1 - 4\delta/h_{\max}q_{\max}$. Thus by a union bound over all qs with a Goldilocks level and summing over these, the 0th level contributes at most

$$\sum_{q\in q_{\max}} \|\mathbf{C}^{(0}\mathbf{A}^{(q)}\mathbf{x}\|_1 \mathbb{1}(\exists h : p_h|W_q| \in [m_{\text{crowd}}, Bm_{\text{crowd}})) \leqslant \sum_{q\in q_{\max}} \frac{\varepsilon}{h_{\max}}\|W_q(\boldsymbol{\lambda})\|_1\|\mathbf{Ax}\|_1 \leqslant \varepsilon\|\mathbf{Ax}\|_1.$$

Let this be event $\mathcal{E}_3$. On the other hand, for the Goldilocks level itself, there is a $1 - 2q_{\max}/\alpha$ probability that

$$\begin{aligned}
&\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]} \|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|\mathbb{1}(\mathcal{E}_{h,q}([m_{\text{crowd}}, Bm_{\text{crowd}}])) \\
\leqslant &\sum_{q\in[q_{\max}]} (1+\varepsilon)\|\mathbf{A}^{(q)}\mathbf{x}\|_1 \mathbb{1}(\exists h : p_h|W_q| \in [m_{\text{crowd}}, Bm_{\text{crowd}})) + 4\varepsilon\|W_q\|_1\|\mathbf{Ax}\|_1 \\
\leqslant &\, 4\varepsilon\|\mathbf{Ax}\|_1 + \sum_{q\in[q_{\max}]} (1+\varepsilon)\|\mathbf{A}^{(q)}\mathbf{x}\|_1 \mathbb{1}(\exists h : p_h|W_q| \in [m_{\text{crowd}}, Bm_{\text{crowd}}))
\end{aligned}$$

by a union bound over the at most $q_{\max}$ weight classes.

Otherwise, if a weight class q has no Goldilocks level, then we have by the triangle inequality that

$$\|\mathbf{C}^{(0}\mathbf{A}^{(q)}\mathbf{x}\|_1 \leqslant \|\mathbf{A}^{(q)}\mathbf{x}\|_1$$

and thus we simply bound the contribution of the 0th level by $\|\mathbf{A}^{(q)}\mathbf{x}\|_1$.

Note that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ occurs with probability at least $1 - 11\delta$. Then conditioned on this event, every $\mathbf{x} \in \mathbb{R}^d$ has a $1 - 2q_{\max}/\alpha$ probability that

$$\begin{aligned}\|\mathbf{SAx}\|_1 &= \left[\sum_{q>q_{\max}} \|\mathbf{SA}^{(q)}\mathbf{x}\|_1\right] + \sum_{q\in[q_{\max}]}\left[\|\mathbf{C}^{(0}\mathbf{A}^{(q)}\mathbf{x}\|_1 + \sum_{h\in[h_{\max}]}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}(\mathbf{A}^{(q)}\mathbf{x}\|_1\right] \\ &\leqslant \varepsilon\|\mathbf{Ax}\|_1 + \underbrace{(1+\varepsilon)\|\mathbf{Ax}\|_1}_{\text{Goldilocks or 0th level}} + \underbrace{\varepsilon\|\mathbf{Ax}\|_1}_{\text{0th level if Goldilocks level exists}} + \underbrace{5\varepsilon\|\mathbf{Ax}\|_1}_{\text{badly concentrated and oversampled levels}} \\ &\leqslant (1+8\varepsilon)\|\mathbf{Ax}\|_1\end{aligned}$$

which is the desired bound. ■

We conclude by a standard net argument.

Theorem B.25 (No expansion) *With probability at least $1 - 11\delta$, we have that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\|\mathbf{SAx}\|_1 \leqslant (1+11\varepsilon)\|\mathbf{Ax}\|_1.$$

Proof By Lemma B.24, there is an event with probability at least $1 - 10\delta$ such that conditioned on this event, for each $\mathbf{x}$, there is a $1 - 2/\alpha$ probability that

$$\|\mathbf{SAx}\|_1 \leqslant (1+8\varepsilon)\|\mathbf{Ax}\|_1. \tag{1}$$

It is well-known (see e.g., Bourgain et al. (1989)), that there exists an ε-net $\mathcal{N}$ of size at most $(3/\varepsilon)^d = \exp(d\log(3/\varepsilon))$ over the set $\{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{Ax}\| = 1\}$. Then by a union bound over the net, Equation 1 holds for every $\mathbf{Ax} \in \mathcal{N}$ with probability at least $1 - \delta$.

Finally, let $\mathbf{x} \in \mathbb{R}^d$ be arbitrary with $\|\mathbf{Ax}\|_1 = 1$. It is shown in (Wang and Woodruff, 2019, Theorem 3.5) that $\mathbf{Ax} = \sum_{i=0}^{\infty}\mathbf{y}^{(i)}$ where each nonzero $\mathbf{y}^{(i)}$ has $\mathbf{y}^{(i)}/\|\mathbf{y}^{(i)}\|_1 \in \mathcal{N}$ and $\|\mathbf{y}^{(i)}\|_1 \leqslant \varepsilon^i$. We then have that

$$\|\mathbf{SAx}\|_1 = \left\|\mathbf{S}\sum_{i=0}^{\infty}\mathbf{y}^{(i)}\right\|_1 \leqslant \sum_{i=0}^{\infty}\left\|\mathbf{Sy}^{(i)}\right\|_1 \leqslant (1+8\varepsilon)\sum_{i=0}^{\infty}\left\|\mathbf{y}^{(i)}\right\|_1 \leqslant (1+8\varepsilon)\sum_{i=0}^{\infty}\varepsilon^i \leqslant 1+11\varepsilon.$$

We conclude by homogeneity. ■

Appendix C. Near Optimal Trade-offs for ℓ_1 Entrywise Embeddings

In this section, we obtain algorithmic trade-offs between sketching dimension and distortion for ℓ_1 entrywise embeddings, and show that this is nearly tight for $d \times d$ matrices.

C.1. Algorithm

Our algorithm is an M-sketch with subsampling rates $p_h = B^{-h}$, where $B = (\frac{d}{\delta}\log n)^{\alpha}$ for $\alpha \in (0,1)$, and CountSketch hashes into $\tilde{\Theta}(\frac{B}{\delta}\log n)$ buckets. By homogeneity, we assume that $\|\mathbf{A}\|_1 = 1$ throughout this section.

Definition C.1 (Useful constants)

$$\begin{aligned} B &:= \left(\frac{d}{\delta}\log n\right)^{\alpha} \\ h_{\max} &:= \log_B n \\ q_{\max} &:= \log_2 \frac{ndh_{\max}}{\delta} \\ p_h &:= B^{-h}, \qquad\qquad h \in [h_{\max}] \end{aligned}$$

Theorem C.2 *Let $\delta \in (0,1)$ and $\alpha \in (0,1)$. Then there exists a sparse oblivious ℓ_1 entrywise embedding $\mathbf{S}$ into k dimensions with*

$$k = \left(\frac{d}{\delta}\log n\right)^{\alpha} \operatorname{poly}(\delta^{-1}, \log n)$$

such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\left\{\Omega(1)\|\mathbf{A}\|_1 \leqslant \|\mathbf{SA}\|_1 \leqslant O\left(\frac{1}{\delta\alpha}\right)\|\mathbf{A}\|_1\right\} \geqslant 1-\delta.$$

Our analysis revolves around the vector of row norms.

Definition C.3 (Row norms vector) *For an $n \times d$ matrix $\mathbf{A}$ with $\|\mathbf{A}\|_1 = 1$, we define the row norms vector $\mathbf{a} \in \mathbb{R}^n$ by $\mathbf{a}_i = \|\mathbf{e}_i^\top \mathbf{A}\|_1$. Using this vector, we define weight classes $W_q(\mathbf{a})$ and restrictions $\mathbf{A}^{(q)}$ of $\mathbf{A}$ to our weight classes, analogously to the analysis in Section B.*

In order to avoid shrinking the vector $\mathbf{a}$ by more than a constant factor with probability at least δ, we apply Lemma G.1 with failure rate δ and constant ε, which gives an M-sketch with 0th level hash bucket size

$$N_0 = \tilde{O}\left(\frac{B}{\delta}\log\log n\right)$$

and hth level hash bucket size

$$N = \tilde{O}(B\log n).$$

We now show that this does not dilate the entrywise 1-norm of $\mathbf{A}$ by more than $O(1/\alpha)$. As in the analysis in Verbin and Zhang (2012), we use the Rademacher dimension.

Lemma C.4 (Rademacher dimension of ℓ_1^d) *Let $\{\mathbf{x}_i\}_{i=1}^s \subseteq \mathbb{R}^d$ with $\|\mathbf{x}_i\|_1 \leqslant 1$ for each $i \in [s]$, and let $\delta \in (0,1)$. Let $\{\varepsilon_i\}_{i=1}^s$ be independent Rademacher variables. Then with probability at least $1-\delta$,*

$$\left\|\sum_{i=1}^s \varepsilon_i \mathbf{x}_i\right\|_1 \leqslant d\sqrt{\frac{1}{2}\log\frac{2d}{\delta}}\sqrt{s}.$$

Proof The proof uses standard concentration inequalities and is similar to (Verbin and Zhang, 2012, Lemma 1). The details are deferred to Appendix H. ■

We follow the approach of Verbin and Zhang (2012). Using the Rademacher dimension, we first show that if we sample too many elements, then the contribution from this level is at most a negligible fraction of the total mass.

Lemma C.5 *Let $q \in [q_{\max}]$. Let $p_h|W_q(\mathbf{a})| \geqslant bN$ for*

$$b = 2d^2 h_{\max}^2 q_{\max}^2 \log\left(\frac{2dNh_{\max}q_{\max}}{\delta}\right)$$

Then with probability at least $1 - \delta/q_{\max}$,

$$\sum_{h:p_h|W_q(\mathbf{a})|\geqslant bN} \|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 \leqslant \frac{1}{q_{\max}}\|\mathbf{A}^{(q)}\|_1$$

Proof By Chernoff bounds, the probability that we sample $(1 \pm 1/2)p_h|W_q(\mathbf{a})|/2N$ elements in a given bucket in the hth level is at most

$$\exp\left(-\frac{(1/2)^2 p_h|W_q(\mathbf{a})|/N}{3}\right) \leqslant \exp\left(-\frac{b}{12}\right) \leqslant \exp\left(-\log\left(\frac{Nh_{\max}q_{\max}}{\delta}\right)\right) = \frac{\delta}{Nh_{\max}q_{\max}}$$

so by a union bound over the N buckets, this holds simultaneously for all buckets at the hth level with probability at least $\delta/h_{\max}q_{\max}$.

We condition on the above event. Then, each bucket L is a randomly signed sum of $s \geqslant b$ elements $\mathbf{e}_i^\top\mathbf{A}$ with $\|\mathbf{e}_i^\top\mathbf{A}\|_1 \leqslant 2^{1-q}$. Thus by Lemma C.4, with probability at least $\delta/Nh_{\max}q_{\max}$,

$$\begin{aligned}\left\|\sum_{\mathbf{a}_i\in L_{h,k}} \Lambda_i \mathbf{e}_i^\top\mathbf{A}\right\|_1 &\leqslant 2^{1-q}d\sqrt{\frac{1}{2}\log\frac{2dNh_{\max}q_{\max}}{\delta}}\sqrt{s} \\ &\leqslant \frac{\|W_q(\mathbf{a})\|_1}{N}\frac{2d}{\sqrt{s}}\sqrt{\frac{1}{2}\log\frac{2dNh_{\max}q_{\max}}{\delta}} \\ &\leqslant \frac{\|W_q(\mathbf{a})\|_1}{Nh_{\max}q_{\max}}\end{aligned}$$

as we have set

$$\frac{2d}{\sqrt{s}}\sqrt{\frac{1}{2}\log\frac{2dNh_{\max}q_{\max}}{\delta}} \leqslant \frac{2d}{\sqrt{b}}\sqrt{\frac{1}{2}\log\frac{2dNh_{\max}q_{\max}}{\delta}} \leqslant \frac{\varepsilon}{h_{\max}q_{\max}}.$$

Summing over the buckets $k \in [N]$ and union bounding and summing over $h \in [h_{\max}]$ yields the desired result. ∎

Next, we handle the remaining levels. We pay the price of having smaller hash buckets in the distortion at this point.

Lemma C.6 *Let $q \in [q_{\max}]$. Then with probability at least $1 - 2\delta/q_{\max}$,*

$$\sum_{h:p_h|W_q(\mathbf{a})|<bN} \|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 \leqslant O\left(\frac{1}{\delta\alpha}\right)\|\mathbf{A}^{(q)}\|_1$$

Proof Note that if $p_h|W_q(\mathbf{a})| \leqslant \delta/h_{\max}q_{\max}$, then by a union bound over the at most $h_{\max}$ levels of h, none of these levels h sample any elements from weight class q with probability at least $\delta/q_{\max}$. Then for each weight class q, only the subsampling levels h for

$$\frac{\delta}{h_{\max}q_{\max}} \leqslant p_h|W_q(\mathbf{a})| \leqslant bN$$

can contribute to the mass of the sketch $\|\mathbf{SA}\|_1$. Note that this is only

$$\log_B\left(\frac{bNh_{\max}q_{\max}}{\delta}\right) = \log_B\left[\mathrm{poly}(d, \log n, \delta^{-1})\right] = O\left(\frac{1}{\alpha}\right)$$

levels of subsampling, where each level contributes at most

$$\mathbf{E}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 \leqslant \mathbf{E}\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 = \|\mathbf{A}^{(q)}\|_1$$

in expectation. We thus conclude by summing over h with $p_h|W_q(\mathbf{a})| < bN$ and then applying a Markov bound. ■

Putting the above pieces together yield the following:

Proof [Proof of Theorem C.2] As previously discussed in this section, the "no contraction" direction of $\|\mathbf{SA}\|_1 \geqslant \Omega(1)\|\mathbf{A}\|_1$ is handled in Lemma G.1, so we focus on proving the "no dilation" direction of $\|\mathbf{SA}\|_1 \geqslant O(1/\delta\alpha)\|\mathbf{A}\|_1$.

We union bound and sum over the results from Lemmas C.5 and C.6 for $q \in [q_{\max}]$ to see that with probability at least $1 - 3\delta$,

$$\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 \leqslant O\left(\frac{1}{\delta\alpha}\right)\sum_{q\in[q_{\max}]}\|\mathbf{A}^{(q)}\|.$$

We also note that $\|\mathbf{C}^{(0}\mathbf{A}\|_1 \leqslant \|\mathbf{A}\|_1$ by the triangle inequality. Finally, we have that in expectation, the weight classes $q > q_{\max}$ contribute at most

$$\begin{aligned}\sum_{q>q_{\max}}\sum_{h\in[h_{\max}]}\mathbf{E}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 &\leqslant \sum_{q>q_{\max}}\sum_{h\in[h_{\max}]}\|\mathbf{A}^{(q)}\|_1\\ &\leqslant \sum_{h\in[h_{\max}]}\frac{2\delta}{ndh_{\max}}\|\mathbf{A}\|_1\sum_{q>q_{\max}}|W_q(\mathbf{a})|\\ &\leqslant 2\delta\|\mathbf{A}\|_1.\end{aligned}$$

Then by Markov's inequality, with probability at least $1 - \delta$, these levels contribute at most $2\|\mathbf{A}\|_1$. Summing these three results, we find that

$$\begin{aligned}\|\mathbf{SA}\|_1 &\leqslant \|\mathbf{C}^{(0}\mathbf{A}\|_1 + \sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1 + \sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\|_1\\ &\leqslant 3\|\mathbf{A}\|_1 + O\left(\frac{1}{\delta\alpha}\right)\sum_{q\in[q_{\max}]}\|\mathbf{A}\|_1\\ &\leqslant O\left(\frac{1}{\delta\alpha}\right)\|\mathbf{A}\|_1\end{aligned}$$

as desired. ■

C.2. Lower bound

We show that for $d \times d$ matrices, the above trade-off between the sketching dimension and distortion is nearly optimal, up to log factors. Note that for constant δ, the above result gives a $d^{\alpha} \operatorname{poly} \log d$ sized sketch with distortion $1/\alpha$. We show that with a sketch of size r, a distortion of $\Omega((\log d)/(\log r))$ is necessary.

By Yao's minimax principle, we assume that the $r \times d$ sketch matrix $\mathbf{S}$ is fixed, and show that the distortion is $\Omega((\log d)/(\log r))$ with constant probability over a distribution over input matrices $\mathbf{A}$.

The following simple lemma is central to our analysis:

Lemma C.7 *Let $\mathbf{S}$ be an $r \times n$ matrix, and let $\mathbf{A}$ be drawn as an $n \times d$ matrix with all of its columns drawn as i.i.d. Cauchy variables. Then,*

$$\mathbf{Pr}\{\Omega(d \log d)\|\mathbf{S}\|_1 \leqslant \|\mathbf{SA}\|_1 \leqslant O(d \log(rd))\|\mathbf{S}\|_1\} \geqslant \frac{99}{100}.$$

Proof The proof relies on standard tricks and is deferred to Appendix H. ■

Theorem C.8 *Let $\mathbf{S}$ be a fixed $r \times d$ matrix. Then there is a distribution μ over $d \times d$ matrices such that if*

$$\Pr_{\mathbf{A} \sim \mu}(\|\mathbf{A}\|_1 \leqslant \|\mathbf{SA}\|_1 \leqslant \kappa \|\mathbf{A}\|_1) \geqslant \frac{2}{3}$$

then $\kappa = \Omega((\log d)/(\log r))$.

Proof We draw our matrix $\mathbf{A}$ from μ as follows. Let μ_1 be the distribution that draws $\mathbf{A}$ as a $d \times d$ i.i.d. matrix with Cauchy entries, and let μ_2 be the distribution that draws $\mathbf{A}$ with its first r columns as a $d \times r$ i.i.d. matrix with Cauchy entries scaled by d/r, and the rest of the $d - r$ columns all 0s. Then, μ draws from μ_1 with probability $1/2$ and μ_2 with probability $1/2$.

Note that by Lemmas 2.10 and 2.12 of Wang and Woodruff (2019),

$$\Pr_{\mathbf{A} \sim \mu_1}\left(\Omega(d^2 \log d) \leqslant \|\mathbf{A}\|_1 \leqslant O(d^2 \log d)\right) \geqslant \frac{99}{100}$$

$$\Pr_{\mathbf{A} \sim \mu_2}\left(\frac{d}{r}\Omega(rd \log(rd)) \leqslant \|\mathbf{A}\|_1 \leqslant \frac{d}{r}O(rd \log(rd))\right) \geqslant \frac{99}{100}$$

By Lemma C.7, if $\mathbf{A} \sim \mu_1$, then $\|\mathbf{SA}\|_1 = \Omega(d \log d)\|\mathbf{S}\|_1$ with probability at least $99/100$. Now suppose for contradiction that $\|\mathbf{S}\|_1 = \omega(\kappa d)$. Then with probability at least $1 - (1/3 + 1/2 + 1/100 + 1/100) > 0$, we have that

$$\omega(\kappa d^2 \log d) = \Omega(d \log d)\|\mathbf{S}\|_1 \leqslant \|\mathbf{SA}\|_1 \leqslant \kappa \|\mathbf{A}\|_1 = O(\kappa d^2 \log d)$$

which is a contradiction. Thus, $\|\mathbf{S}\|_1 = O(\kappa d)$.

Now consider $\mathbf{A} \sim \mu_2$. By Lemma C.7, $\|\mathbf{SA}\|_1 \leqslant O(r \log r)\|\mathbf{S}\|_1$ with probability at least $99/100$. Then, with probability at least $1 - (1/3 + 1/2 + 1/100 + 1/100) > 0$,

$$\Omega(d^2 \log d) \leqslant \|\mathbf{A}\|_1 \leqslant \|\mathbf{SA}\|_1 \leqslant O(d \log r)\|\mathbf{S}\|_1 = O(\kappa d^2 \log r)$$

so

$$\kappa = \Omega\left(\frac{\log d}{\log r}\right)$$

as desired. ■

Appendix D. Independence Testing in the ℓ_1 norm

In this section, we present our result for estimating $\|P - Q\|_1$, where P is the joint distribution and Q the product distribution defined by the marginals, which are determined by the stream items as introduced in Section 1. We first prepare a heavy hitter data structure in Section D.1 and present our $(1+\varepsilon)$-approximation algorithm to the ℓ_1 norm of order-2 tensors in Section D.2. To move to higher dimensions, we need a rough estimator for the product distribution in Section D.3. Finally, we apply the result for order-2 tensors iteratively in Section D.4 to obtain a $(1+\varepsilon)$-approximation to $\|P - Q\|_1$.

D.1. Heavy Hitters

This subsection is devoted to a data structure, called the HEAVYHITTER structure, which is analogous to the classical CountSketch data structure for a general functional f on a general linear space.

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is function satisfying the following properties:

1. $f(0) = 0$;
2. $f(x) = f(-x)$;
3. $f(x)$ is increasing on $[0, \infty)$;
4. There exists a constant C_f such that it holds for any integer $s \geqslant 1$ and any $x_1, \dots, x_s, y_1, \dots, y_s \in \mathbb{R}$ that $\sum_{i=1}^s f(x_i + y_i) \leqslant C_f(\sum_{i=1}^s (f(x_i) + f(y_i)))$.
5. There exists a function $h : [0, \infty) \to [0, \infty)$ such that
   (a) $\lim_{\varepsilon \to 0^+} h(\varepsilon) = 0$;
   (b) it holds for any integer $s \geqslant 1$ and any $x_1, \dots, x_s, y_1, \dots, y_s \in \mathbb{R}$ that $|\sum_{i=1}^s f(x_i + y_i) - \sum_{i=1}^s f(x_i)| \leqslant h(\varepsilon) \sum_i f(x_i)$ whenever $\sum_i f(y_i) \leqslant \varepsilon \sum_i f(x_i)$.

We abuse notation and define for $x \in \mathbb{R}^m$ that $f(x) = \sum_i f(x_i)$.

We define a different Rademacher dimension as follows. The Rademacher dimension $B = B(f; \eta)$ is the smallest integer such that the following holds for any integer $s \geqslant 1$. Let $\sigma_1, \dots, \sigma_s$ be i.i.d. Rademacher variables and $\xi_1, \dots, \xi_s$ be i.i.d. Bernoulli variables such that $\mathbf{E}\, \xi_i = 1/B$. It holds for any $x_1, \dots, x_s \in \mathbb{R}^m$ that

$$\mathbf{Pr}\left\{ f\left(\sum_i \sigma_i \xi_i x_i \right) \leqslant \eta \sum_i f(x_i) \right\} \geqslant 0.9.$$

Lemma 1 *Let $\gamma, \zeta \in (0, 1/3)$, there exists $r = r(\gamma, \zeta)$ and a randomized linear map $T : \mathbb{R}^m \to \mathbb{R}^r$, and a subrecovery algorithm $\mathcal{B}$ such that for each $x \in \mathbb{R}^m$, with probability at least $1 - \zeta$, it holds that $(1-\gamma)f(x) \leqslant \mathcal{B}(Tx) \leqslant (1+\gamma)f(x)$.*

Then, for $\theta, \delta \in (0, 1/3)$, there exists a randomized linear function $M : (\mathbb{R}^m)^d \to \mathbb{R}^S$, where $S = O(B \log(d/\delta) \cdot r(\gamma, \zeta'))$ for $B = B(f; h^{-1}(\theta)\theta)$ and $\zeta' = O(\zeta/(B \log(d/\delta)))$, and a recovery algorithm $\mathcal{A}$ satisfying the following. For any $x = (x_1, \dots, x_n) \in (\mathbb{R}^m)^d$ with probability $\geqslant 1 - \delta - \zeta$, $\mathcal{A}$ reads Mx and outputs an estimate $\tilde{f}_i$ for each $i \in [d]$ such that

1. *$|\tilde{f}_i - f(x_i)| \leqslant (\gamma + \theta + \gamma\theta) f(x_i)$ whenever $f(x_i) \geqslant \theta f(x)$;*
2. *$|\tilde{f}_i| \leqslant C_f \theta (1+\gamma)(1 + h(\theta)) f(x)$ whenever $f(x_i) < \theta f(x)$.*

Proof The linear sketch M is essentially a CountSketch data structure, which hashes $\{Tx_i\}_i$ into $B = B(f; \min\{\varepsilon^2, h(\theta)\}\eta)$ buckets under a hash function h. The b-th bucket contains

$$T_b = \sum_{i:h(i)=b} \sigma_i T x_i.$$

For i^* such that $f(x_{i^*}) \geqslant \theta f(x)$, the algorithm will just return $\tilde{f}_{i^*} = \mathcal{B}(T_{h(i^*)})$. Next we analyse the estimation error. Let $b = h(i^*)$. Note that $\sum_{i:h(i)=b} \sigma_i T x_i$ is identically distributed as $Tx_{i^*} + T\nu$, where $\nu = \sum_{i \neq i^*: h(i)=b} \sigma_i x_i$. Since $B = B(f; h^{-1}(\theta)\theta)$, it holds with probability at least 0.9 that

$$f(\nu) \leqslant h^{-1}(\theta)\theta f(x) \leqslant h^{-1}(\theta) f(x_{i^*}),$$

which implies that

$$(1-\theta)\, f(x_{i^*}) \leqslant f\left(x_{i^*} + \nu\right) \leqslant (1+\theta)\, f(x_{i^*}).$$

and, with probability at least $0.9 - \zeta$ that

$$(1-\gamma)(1-\theta)\mathcal{B}(Tx_{i^*}) \leqslant \mathcal{B}(T_b) \leqslant (1+\gamma)(1+\theta)\mathcal{B}(Tx_{i^*}).$$

On the other hand, when $f(x_{i^*}) \leqslant \theta f(x)$,

$$f(x_{i^*} + \nu) \leqslant C_f(f(x_{i^*}) + f(\nu)) \leqslant C_f(\theta + h(\theta)\theta) f(x)$$

and, with probability at least $1 - \zeta$,

$$\mathcal{B}(T_b) \leqslant C_f \theta (1+\gamma)(1 + h(\theta)) f(x).$$

Repeat $\Theta(\log(d/\delta))$ times to drive the failure probability down to δ/d to take a union bound over all i^*. ■

The data structure described in Lemma 1 is our HeavyHitter structure, parameterized with (θ, δ).

D.2. $(1+\varepsilon)$-Approximator

Suppose that for any $\gamma, \zeta \in (0,1)$ that are small enough, there exist $t = t(\gamma, \zeta)$, a randomized linear map $T : \mathbb{R}^m \to \mathbb{R}^t$ and a subrecovery algorithm $\mathcal{B}$ such that for each $x \in X$, with probability at least $1 - \zeta$, it holds that $(1-\gamma)f(x) \leqslant \mathcal{B}(Tx) \leqslant (1+\gamma)f(x)$.

Let $x = (x_1, \ldots, x_N) \in (\mathbb{R}^m)^N$. In this subsection, we consider the problem of approximating $M = \sum_i f(x_i)$ up to a $(1+\varepsilon)$-factor. We also assume that we have an approximation $\widehat{M}$ to M such that $M \leqslant \widehat{M} \leqslant KM$.

Our algorithm is inspired from arguments in Andoni et al. (2009). We prepare the following data structure (Algorithm 1) with the entry update algorithm (Algorithm 2). The recovery algorithm is presented in Algorithm 3.

Theorem 2 *Let $\varepsilon \in (0,1)$ be small enough and $K \geqslant 2$ be a power of 2. Let θ, B, Q be as defined in Algorithm 1. There exists an absolute constant $\alpha < 1$ and a randomized linear sketch $\Pi : (\mathbb{R}^m)^N \to \mathbb{R}^S$, where $S = O(Q \cdot t(\alpha\varepsilon/2, 0.05/Q))$ and a recovery algorithm $\mathcal{A}$ satisfying the following.*

For any $x = (x_1, \ldots, x_N) \in X^N$ and an approximation $\widehat{M} \geqslant M = \sum_i f(x_i)$, with probability at least 0.6, $\mathcal{A}$ reads Πx and outputs $\widetilde{M}(x)$ such that

Algorithm 1: Data Structure for constant failure probability algorithm (SubsamplingHeavyHitter)

Require: $\varepsilon, \delta, K, N, t, \zeta$

$L \leftarrow \log(KN/\varepsilon)$

$\hat{L} \leftarrow \log N$

$\theta \leftarrow \min\{\Theta(\varepsilon^3/(C_f L^3)), h^{-1}(\alpha\varepsilon/3), \alpha\varepsilon/4\}$

$B \leftarrow B(f; h^{-1}(\theta)\theta\})$

$Q \leftarrow O(B(\hat{L}+1)\log(N\hat{L}))$

Instantiate a subsampling function H, which hashes $[N]$ into $\hat{L}$ levels such that the sampling probability for the ℓ-th level is $2^{-\ell}$ and is pairwise independent

for *each* $\ell = 0, 1, \ldots, \hat{L}$ **do**

  Instantiate a HeavyHitter structure $\mathcal{D}_\ell$ with parameters $(\theta, 0.05/(\hat{L}+1))$, in which each bucket stores a vector of length $t = t(\alpha\varepsilon/2, \zeta/Q)$

end

Algorithm 2: Algorithm for an update to x_i for our constant failure probability algorithm

Input: an update of the form $x_i \leftarrow x_i + \Delta x_i$

for *each* $\ell = 0, \ldots, \hat{L}$ **do**

  if H *hashes* i *into level* ℓ **then** ▷ Assume H hashes every i to level 0

    $b_\ell \leftarrow$ index of the bucket containing i in $\mathcal{D}_\ell$

    Add $T(\Delta x_i)$ to the b_ℓ-th bucket

  end

end

Algorithm 3: $(1+\varepsilon)$-approximator to $f(x)$ with constant failure probability

Require: (i) A subsampling scheme H such that the i-th level has subsampling probability $p_i = 2^{-i}$; (ii) $\hat{L}+1$ HeavyHitter structures $\mathcal{D}_0, \dots, \mathcal{D}_{\hat{L}}$ with the same parameters (θ, δ), where $\hat{L} = \log N$, $\theta = \min\{\Theta(\frac{\varepsilon^3}{C_f L^3}), \frac{\alpha\varepsilon}{4}, h^{-1}(\frac{\alpha\varepsilon}{3})\}$ and $\delta = \frac{1}{20(\hat{L}+1)}$; (iii) an approximation $\widehat{M}$ such that $M \leqslant \widehat{M}$; (iv) an integer $K \geqslant 2$ which is a power of 2.

$L \leftarrow \log(2N/\varepsilon)$
$\hat{L} \leftarrow \log N$
for $j = 0, \dots, \hat{L}$ **do**
  $\Lambda_j \leftarrow$ top $\Theta(L^3/\varepsilon^3)$ heavy hitters from $\mathcal{D}_j$
end
$j_0 \leftarrow \log(4K\varepsilon^{-3}L^3)$
$\zeta \leftarrow$ uniform variable in $[1/2, 1]$
for $j = 0, \dots, j_0$ **do**
  Let $\lambda_1^{(j)}, \dots, \lambda_s^{(j)}$ be the elements in Λ_0 contained in $[(1+\varepsilon)\zeta\frac{\widehat{M}}{2^j}, (2-\varepsilon)\zeta\frac{\widehat{M}}{2^j}]$
  $\widetilde{M}_j \leftarrow f(\lambda_1^{(j)}) + \cdots + f(\lambda_s^{(j)})$
end
for $j = j_0+1, \dots, L$ **do**
  Find the biggest ℓ such that Λ_ℓ contains s elements $\lambda_1^{(j)}, \dots, \lambda_s^{(j)}$ in $[(1+\varepsilon)\zeta\frac{\widehat{M}}{2^j}, (2-\varepsilon)\zeta\frac{\widehat{M}}{2^j}]$ for $(1-\sqrt{20}\varepsilon)\frac{L^2}{\varepsilon^2} \leqslant s \leqslant 2(1+\sqrt{20}\varepsilon)\frac{L^2}{\varepsilon^2}$
  if *such ℓ exists* **then**
    $\widetilde{M}_j \leftarrow (f(\lambda_1^{(j)}) + \cdots + f(\lambda_s^{(j)}))2^\ell$
  else
    $\widetilde{M}_j \leftarrow 0$
  end
end
return $\widetilde{M} \leftarrow \sum_j \widetilde{M}_j$

1. $(1-\varepsilon)M \leqslant \widetilde{M}(x) \leqslant (1+\varepsilon)M$ *if* $\widehat{M} \in [(K/2)M, KM]$;

2. $\widetilde{M}(x) \leqslant (1+\varepsilon)M$ *otherwise.*

Proof There are $\Theta(\log(1/\delta))$ repetitions. In each repetition, there are $(\hat{L}+1)$ HeavyHitter structures of $O(B\log(N\hat{L}))$ buckets. There are $O(B(\hat{L}+1)\log(N\hat{L}))$ buckets in each repetition. Each bucket stores a sketch of length $t(\alpha\varepsilon/2, 0.05/Q)$. The total space complexity follows.

Since for each bucket the failure probability is $0.05/Q$, we can take a union bound over all buckets and assume that $\mathcal{B}$ gives accurate answers on all buckets in a repetition with probability at least 0.95. Then the claimed result follows from Theorem 6 for $\hat{M} \in [(K/2)M, KM]$ and from Lemma 11 for $\widehat{M} > KM$ and Lemma 12 for $\widehat{M} < (K/2)M$. ■

Next we extend the algorithm to handle the case where $\widehat{M} < (K/2)M$.

Theorem 3 *Let $\varepsilon, \theta, B, Q, S$ be as in Theorem 2 and $\delta \in (0,1)$. There exists an absolute constant $\alpha < 1$ and a randomized linear sketch $\Pi : (\mathbb{R}^m)^N \to \mathbb{R}^{S'}$, where $S' = O(S\log K \cdot \log(\delta^{-1}\log K))$ and a recovery algorithm $\mathcal{A}$ satisfying the following.*

For any $x = (x_1, \ldots, x_N) \in X^N$ and an approximation $\widehat{M}$ such that $M \leqslant \widehat{M} \leqslant KM$, where $M = \sum_i f(x_i)$, with probability at least $1-\delta$, $\mathcal{A}$ reads Πx and outputs $\widetilde{M}(x)$ such that $(1-\varepsilon)M \leqslant \widetilde{M}(x) \leqslant (1+\varepsilon)M$.

Proof First, in view of Theorem 2, repeating the Algorithm 3 $\Theta(\log(1/\zeta))$ times and taking the median reduces the failure probability of a single run to ζ. Hence, with sketch length $O(S\log(1/\zeta))$, we have an algorithm outputting $\widetilde{M}$ such that $(1-\varepsilon)M \leqslant \widetilde{M}(x) \leqslant (1+\varepsilon)M$, provided that $(K/2)M \leqslant \widehat{M} \leqslant KM$.

For a general $\widehat{M}$, we run $\log K$ instances of the aforesaid algorithm in parallel, where the parameter K in Algorithm 3 takes values $2, 4, 8, \ldots, K$, respectively. Note that $\widehat{M} \in [(K/2)M, KM]$ in one of these instances and, with probability at least $1-\zeta$, the output $\widehat{M}$ of this instance satisfies that $\widehat{M} \in [(1-\varepsilon)M, (1+\varepsilon)M]$. For each other instance, with probability at $1-\zeta$, the outputted $\widehat{M} \leqslant (1+\varepsilon)M$. Setting $\zeta = \delta/(\log K)$ and taking the maximum output $\widehat{M}$ among the $\log K$ instances with a union bound over $\log K$ instances, we obtain an estimate in $[(1-\varepsilon)M, (1+\varepsilon)M]$ with probability at least $1-\delta$, as desired. ■

D.3. Rough Approximator for ℓ_1-Norm

Consider the problem of estimating $\|x\|_1$ up to a constant factor for $x \in \mathbb{Z}^{d^q}$ in the turnstile streaming model, where each update changes a coordinate by a $+1$ or a -1. Let $N = d^q$. The following result is due to Braverman and Ostrovsky Braverman and Ostrovsky (2009).

Theorem 4 (Rough approximation; Corollary 6.6 and Lemma 6.7 in Braverman and Ostrovsky (2009)) *There exists a randomized linear sketch $\Pi : \mathbb{Z}^N \to \mathbb{Z}^S$ for $S = \tilde{O}(q\log(md))$ and a recovery algorithm $\mathcal{A}$ satisfying the following. For any $x \in \mathbb{Z}^N$ given in the aforementioned turnstile streaming model of length m, with probability at least 0.95, $\mathcal{A}$ reads Πx and outputs $\widehat{M}$ such that $\|x\|_1 \leqslant \widehat{M} \leqslant 4^{q^2}(\log d)^q \|x\|_1$.*

D.4. Estimation of Total Variation Distance

Now we wish to estimate $\|P - Q\|_1$. Recall that P is a general joint distribution and Q the product distribution induced by the marginals of P.

We apply the data structure iteratively in Section D.2. For ℓ_1-norm, $f(x) = h(x) = x$, $C_f = 1$, $B(f;\varepsilon) = \Theta(1/\varepsilon^2)$. Therefore, in Theorem 3, one can take $B_i = (L/\varepsilon)^c$ for some absolute constant $c \geqslant 4$. The basic setup is presented in Algorithm 4. For each i, we apply Theorem 3 and obtain a linear sketch $\Pi^{(i)}$ and a recovery algorithm $\mathcal{A}_i$. The sub-recovery algorithm for $\mathcal{D}_\ell^{(i)}$ is $\mathcal{A}_{i-1}$. The entry update calls $\texttt{EntryUpdate}(i_1, \ldots, i_q, \Delta, q)$ on the final sketch (see Algorithm 5) if there is an entry update of Δ at position $(i_1, \ldots, i_q)$. For notational convenience, we assume it is always true that a subsampling hash function hashes all coordinates into level 0. The overall decoding algorithm calls $\texttt{Decode}(q)$, see Algorithm 6.

Algorithm 4: Data Structure for P

$\varepsilon_q \leftarrow \varepsilon$, $\delta_q \leftarrow \delta$, $K \leftarrow 4^{q^2} \log^q d$, $N_q \leftarrow d$, $L_q \leftarrow \log(KN_q/\varepsilon_q)$
for *each* $i = q-1, \ldots, 1$ **do**
  $\varepsilon_i \leftarrow \alpha\varepsilon_{i+1}$
  $N_i \leftarrow d$
  $L_i \leftarrow \log(KN_i/\varepsilon_i)$
  $\delta_i \leftarrow O(1/(L_i/\varepsilon_i)^c)$
end
$t_0 \leftarrow 1$
for *each* $i = 1, \ldots, q$ **do**
  $t_i \leftarrow O((L_i/\varepsilon_i)^c t_{i-1} \log K \log(K/\delta_i))$ ▷ sketch length in each bucket
1 $R_i \leftarrow \Theta(\log(1/\delta_i))$ ▷ number of repetitions
2 **for** *each* $r = 1, \ldots, R_i$ **do**
3 Initialize $H^{(i,r)}, \mathcal{D}_0^{(i,r)}, \ldots, \mathcal{D}_{\log N+1}^{(i,r)}$ as in Algorithm 1 for parameters $\varepsilon_i, \delta_i, K, N_i, t_{i-1}$
4 **end**
end

Theorem 5 *Suppose that the stream length* $m = \mathrm{poly}(d^q)$. *There is a randomized sketching algorithm which outputs a* $(1 \pm \varepsilon)$*-approximation to* $\|P - Q\|_1$ *with probability at least* 0.9*, using* $\exp(O(q^2 + q\log(q/\varepsilon) + q\log\log d))$ *bits of space. The update time is* $\exp(O(q^2 + q\log(q/\varepsilon) + q\log\log d))$.

Proof Let P^f be the frequency vector of the empirical distribution of the input stream and P_i^f be the corresponding frequency vector for the marginal on X_i. We have $P = P^f/m$ and $P_i = P_i^f/m$.

Let $\Pi^{(q)}$ be the final linear sketch described above. In parallel we run the rough approximator (Theorem 4), which applies in our setting because the stream items are samples from the distribution and we are counting the empirical frequency. We maintain $\Pi(P^f)$ as described in Algorithm 2. For the marginals P_i^f, we maintain sketches $S_0^{(i)} P_i^f, \ldots, S_{L'}^{(i)} P_i^f$ as in Algorithm 7 and Algorithm 8. At the end of the stream, we construct $\Pi^{(q)} Q^f$ for $Q^f = P_1^f \otimes \cdots \otimes P_n^f$ as in Algorithm 9. Then we

Algorithm 5: Update algorithm for P: an entry update of Δ at position $(i_1, \ldots, i_q)$

Function `EntryUpdate` $(i_1, \ldots, i_q, \Delta, d)$ **:** ▷ invoked on some sketch structure

for *each pair* $(r, \ell) \in [R_d] \times \{0, \ldots, \log N_d\}$ **do**
  if $H^{(d,r)}$ *hashes* i_d *into the* ℓ*-th level* **then**
    $B \leftarrow$ set of indices of buckets containing i_d in $\mathcal{D}_\ell^{(d,r)}$
    for *each bucket* $b \in B$ **do**
      if $d > 1$ **then**
        $\Delta' \leftarrow$ `EntryUpdate` $(i_1, \ldots, i_q, \Delta, d-1)$ on bucket b
        Add Δ' to b
      else
        Add $T(\Delta)$ to b
      end
    end
  end
end
return the incremental vector to the sketch under $\Pi^{(d)}$

Algorithm 6: Decoding algorithm $\mathcal{A}_d$ (for P and $P - Q$)

Function `Decode` (d) **:** ▷ This is $\mathcal{A}_d$

for *each* $r = 1, \ldots, R_d$ **do**
  $Z_r \leftarrow$ Output of Algorithm 3 with subdecoding algorithm $\mathcal{A}_{d-1}$
end
return $\text{median}_r\, Z_r$

Algorithm 7: Data Structure for Q

Let $\varepsilon_i, \delta_i, K, R_i$ be the same as in Algorithm 4

Let the HeavyHitter sketches $\hat{\mathcal{D}}_\ell^{(i,r)}$ be the same as $\mathcal{D}_\ell^{(i,r)}$ in Algorithm 1 (same hash functions) except for $t_i = 1$

Algorithm 8: Entry update for Q

Input: an update of Δ at position $(i_1, \ldots, i_q)$

for *each* $d = 1, \ldots, q$ **do**
  for *each* $(r, \ell) \in [R_d] \times \{0, \ldots, \log N_d\}$ **do**
    if $H^{(d,r)}$ *hashes* i_d *in level* ℓ **then**
      $B \leftarrow$ set of indices of buckets containing i_d in $\hat{\mathcal{D}}_\ell^{(j,r)}$
      Add Δ to bucket b for every $b \in B$
    end
  end
end

Algorithm 9: Tensorization of Q: construct the sketch for $P_1^f \otimes \cdots \otimes P_n^f$

$v^{(0)} = 1$
for *each* $i = 1, \ldots, d$ **do**
  for *each* $(r, \ell) \in [R_d] \times \{0, \ldots, \log N_d\}$ **do**
    if $H^{(d,r)}$ *hashes* i_d *in level* ℓ **then**
      $B \leftarrow$ set of indices of buckets containing i_d in $\mathcal{D}_\ell^{(j,r)}$
      for *each* $b \in B$ **do**
        $a \leftarrow$ bucket value of bucket b in $\hat{\mathcal{D}}_\ell^{(j,r)}$
        Add $a \cdot v^{(i-1)}$ to bucket b
      end
    end
  end
  Form $v^{(i)}$ which conforms to the structure of $\Pi^{(i)}$
end

compute $m^{q-1}\Pi^{(q)}P^f - \Pi^{(q)}Q^f = m^q\Pi(P-Q)$, from which we can recover an approximation to $\|P-Q\|_1$ by invoking $\mathcal{A}_q$.

Next we analyze the space complexity. Let $N_i = d$. Since we are sketching $m^q(P-Q)$, whose ℓ_1 norm is an integer and is at most $2m^q$, we see that $K \leqslant 2m^q = d^{\Theta(q^2)}$ by our assumption that $m = \mathrm{poly}(d^q)$. Set $\varepsilon_q = \varepsilon$ and $\delta_q = O(1)$, then

$$\varepsilon_{i-1} = \alpha\varepsilon_i, \quad \delta_{i-1} = \mathrm{poly}\left(\frac{\varepsilon_i}{\log(KN_i/\varepsilon_i)}\right)$$

for all i. This implies that

$$\varepsilon_{q-i} = \alpha^{i-1}\varepsilon, \quad \delta_{q-i} = \frac{\varepsilon^{\Theta(i)}}{q^{\Theta(i)}\log^{\Theta(i)}(\frac{qd}{\varepsilon})}$$

Therefore the target dimension of $\Pi^{(i)}$ is

$$\begin{aligned} t_{i+1} &\leqslant C\left(\frac{\log(KN_i/\varepsilon_{i+1})}{\varepsilon_{i+1}}\right)^c \cdot t_i \cdot \log K \log\left(\frac{K}{\delta_{i+1}}\right) \\ &\leqslant C'\left(\frac{q^2}{\alpha^{q-i}\varepsilon}\right)^{\Theta(1)} \cdot t_i \cdot (q-i)\,\mathrm{polylog}\left(\frac{qd}{\varepsilon}\right) \end{aligned}$$

with $t_1 = 1$. This implies that

$$t_n \leqslant (C')^q \frac{q^{\Theta(q)}}{\alpha^{\Theta(q^2)}\varepsilon^{\Theta(q)}} \cdot q! \cdot \log^{O(q)}\left(\frac{qd}{\varepsilon}\right) = \exp\left(O\left(q^2 + q\log\frac{q}{\varepsilon} + q\log\log d\right)\right).$$

This space dominates the space needed by the rough estimator. Each coordinate requires $O(\log(m^q)) = O(q^2\log d)$ bits and the overall space complexity (in bits) follows.

The update time is clearly dominated by the update time for P, which is dominated by the sketch length. ∎

D.5. Correctness of Algorithm 3

We adopt the notation from Section D.2. Recall that our goal is to estimate $M = \sum_i f(x_i)$ up to a $(1+\varepsilon)$-factor and we also assume that we have an approximation $\widehat{M}$ to M which satisfies that $(K/2)M \leqslant \widehat{M} \leqslant KM$.

Let ζ be a uniform random variable on $[1/2, 1]$. For a magnitude level j, define

$$T_j = \zeta \frac{\widehat{M}}{2^j},$$

and

$$S_j = \{i \in U : f(x_i) \in (T_j, 2T_j]\}, \quad s_j = |S_j|.$$

Observe that if we scale K by a factor of 2^t, the magnitude levels are shifted by t levels (new top levels are empty). It is easy to see that the behaviour of Algorithm 3 is invariant under the concurrent scaling of K and shifting of the magnitude levels (since the bucket contents in the HeavyHitter structures remain the same), we may, with loss of generality, assume that $K = 2$ and $M \leqslant \widehat{M} \leqslant 2M$.

Observe that $\sum_{j \geqslant 1} \sum_{i \in S_j} f(x_i) = M$. Note that each element in level $j > \log(2N/\varepsilon)$ is at most $\widehat{M}/2^j < (\varepsilon/(2N))\widehat{M} < (\varepsilon/N)M$, so it contribute at most εM and thus can be omitted. That is, we only need to consider the levels up to $L = \log(2N/\varepsilon)$.

We call a level j important if

$$\frac{s_j}{2^j} \geqslant \frac{\varepsilon}{2L}$$

and we let $\mathcal{J}$ denote the set of important levels j. The non-important levels contributes at most

$$\sum_{j \notin \mathcal{J}} \sum_{i \in S_j} f(x_i) \leqslant \sum_{j \notin \mathcal{J}} \frac{2\varepsilon}{2L} M \leqslant \varepsilon \widehat{M} \leqslant 2\varepsilon M.$$

The goal of this section is to prove the following theorem.

Theorem 6 *Algorithm 3 returns an estimate $\widetilde{M}$, which with probability at least 0.7 (over ζ and subsampling) satisfies that*

$$(1 - O(\varepsilon))M \leqslant \widetilde{M} \leqslant (1 + O(\varepsilon))M.$$

The rest of the section is devoted to the proof of the theorem. We assume that all COUNT-MIN structures return correct values, at the loss of 0.05 probability. The main argument is decomposed into the following lemmas.

Lemma 7 *With probability at least 0.95 (over subsampling), the following holds for all $j > j_0$ and $j \in \mathcal{J}$. There exists an ℓ such that the substream induced by H_ℓ contains at least $(1 - O(\varepsilon))L^2/\varepsilon^2$ and at most $2(1 + O(\varepsilon))L^2/\varepsilon^2$ elements of S_j. Furthermore, it holds $\frac{3}{4}L^2/\varepsilon^2 \leqslant s_j 2^{-\ell} \leqslant \frac{9}{4}L^2/\varepsilon^2$ for any such ℓ.*

Proof Since $j \in \mathcal{J}$ and $j > j_0$,

$$s_j \geqslant \frac{\varepsilon}{2L} \cdot 2^{j_0} = \frac{4}{\varepsilon^2} L^2.$$

Let $N_{j,\ell}$ denote the number of survivors in the ℓ-th subsampling level. For $\ell = 1$, we have

$$\mathbf{E}\, N_{j,\ell} = s_j 2^{-\ell} \geqslant \frac{4L^2}{\varepsilon^2} 2^{-\ell} = \frac{2L^2}{\varepsilon^2}.$$

Note that $s_j \leqslant 2^j$, and thus for $\ell = j - \log(\varepsilon^{-2}L^2) > 0$, we have

$$\mathbf{E}\, N_{j,\ell} = s_j 2^{-\ell} \leqslant 2^{j-\ell} = \varepsilon^{-2}L^2$$

survivors after sampling. Hence, there exists ℓ such that $\varepsilon^{-2}L^2 \leqslant \mathbf{E}\, N_{j,\ell} \leqslant 2\varepsilon^{-2}L^2$. For any such ℓ, since H_ℓ is pairwise independent, we have $\mathbf{Var}(N_{j,\ell}) \leqslant \mathbf{E}\, N_{j,\ell}$ and it follows from Chebyshev's inequality that with probability at least $1 - 1/(20L^2)$,

$$N_{j,\ell} = \mathbf{E}\, N_{j,\ell} \pm \sqrt{20L^2 \cdot \mathbf{E}\, N_{j,\ell}},$$

that is,

$$(1 - \sqrt{20}\varepsilon)\frac{L^2}{\varepsilon^2} \leqslant N_{j,\ell} \leqslant 2(1 + \sqrt{20}\varepsilon)\frac{L^2}{\varepsilon^2}. \tag{2}$$

A similar argument shows that for each ℓ, with probability at least $1 - 1/(20L^2)$, we have $N_{j,\ell} \geqslant \frac{9}{4}(1-\sqrt{20}\varepsilon)L^2/\varepsilon^2$ if $\mathbf{E}\, N_{j,\ell} \geqslant \frac{9}{4}L^2/\varepsilon^2$ and $N_{j,\ell} \leqslant \frac{3}{4}(1+\sqrt{20}\varepsilon)L^2/\varepsilon^2$ if $\mathbf{E}\, N_{j,\ell} \leqslant \frac{3}{4}L^2/\varepsilon^2$. Taking a union bound over all L, we have that with probability at least $1 - 1/(20L)$ there exists a unique ℓ such that (2) holds; furthermore, $s_j 2^{-\ell} = \mathbf{E}\, N_{j,\ell} = \Theta(\varepsilon^{-2}L^2)$ for this ℓ.

The claimed result follows from a union bound over j. ■

Let $\alpha \in (0, 1)$ be a small constant. Define

$$S_j^* = \{i \in S_j : f(x_i) \in [(1 + (1-\alpha)\varepsilon)T_j, (2 - (1-\alpha)\varepsilon)T_j]\}$$
$$S_j^{**} = \{i \in S_j : f(x_i) \in [(1 + (1+\alpha)\varepsilon)T_j, (2 - (1+\alpha)\varepsilon)T_j]\}$$

and

$$M_j^* = \sum_{i \in S_j^*} f(x_i), \quad M_j^{**} = \sum_{i \in S_j^{**}} f(x_i).$$

Suppose that the event in Lemma 7 occurs.

Lemma 8 *With probability at least* 0.9 *(over the subsamplings), it holds for each $j \in \mathcal{J}$ and $j > j_0$ that $(1 - O(\eta))M_j \leqslant \mathbf{E}\, \widetilde{M}_j \leqslant (1 + O(\eta))$, where the expectation is taken over the subsampling.*

Proof Let $I_\ell \subseteq [N]$ be the set of indices in subsampling level ℓ, where ℓ is found in Step 13 of Algorithm 3. Then

$$\mathbf{E}\, f(x_{I_\ell}) = \frac{f(x)}{2^\ell}.$$

By Lemma 7 and our choice of β, we have

$$\frac{s_j}{2^\ell} \leqslant \frac{9}{4} \cdot \frac{L^2}{\varepsilon^2}. \tag{3}$$

Together with the assumption that $j \in \mathcal{J}$,

$$2^j \leqslant \frac{L}{\varepsilon} s_j \leqslant \frac{9}{4} \cdot \frac{L^3}{\varepsilon^3} 2^\ell,$$

which implies that (by adjusting constants)

$$\frac{\varepsilon^3}{L^3} \mathbf{E}\, f(x_{I_\ell}) \leqslant \frac{9}{4} \frac{f(x)}{2^j} = \frac{9}{4} \frac{1}{\zeta} T_j \leqslant \frac{9}{2} T_j.$$

Except with probability $0.05/L$, we have

$$\frac{\varepsilon^3}{180 L^3} f(x_{I_\ell}) \leqslant T_j.$$

Now, let $\theta = \min\{\varepsilon^3/(180 C_f L^3), \alpha\varepsilon/4, h^{-1}(\alpha\varepsilon/3)\}$ in Lemma 1, we have the guarantees that (1) if $f(x_i) \geqslant \theta T_j$ then it is estimated up to an additive error of at most

$$(\gamma + \theta + \gamma\theta) f(x_i) \leqslant (\gamma + 2\theta) f(x_i) \leqslant \left(\frac{\alpha\varepsilon}{2} + 2 \cdot \frac{\alpha\varepsilon}{4}\right) f(x_i) = \alpha\varepsilon f(x_i),$$

and (2) if $f(x_i) \leqslant \theta T_j$ we obtain an estimate at most

$$\begin{aligned} C_f \theta (1+\gamma)(1+h(\theta)) f(x_{I_\ell}) \leqslant \frac{\varepsilon^3}{40 L^3} (1+\gamma)(1+h(\theta)) f(x_{I_\ell}) &\leqslant (1+\gamma)(1+h(\theta)) T_j \\ &\leqslant \left(1 + \frac{\alpha\varepsilon}{2}\right)\left(1 + \frac{\alpha\varepsilon}{3}\right) T_j \\ &\leqslant (1+\alpha\varepsilon) T_j. \end{aligned}$$

Hence, all survivors in level S_j^{**} will be recovered and all survivors in the higher levels will not be mistakenly recovered in level j; survivors from lower levels will not collude to form a heavy hitter.

Let $R^{(j)} = \{i_1, \ldots, i_s\}$, we have $\lambda_r^{(j)} = (1 + O(\varepsilon)) f(x_{i_r})$ for all $r \in [s]$. Then

$$\widetilde{M}_j = (\lambda_1^{(j)} + \cdots + \lambda_s^{(j)}) 2^\ell = (1 \pm (O(\varepsilon)) \widetilde{M}'_j,$$

where

$$\widetilde{M}'_j = 2^\ell \sum_{r=1}^{s} f(x_{i_r})$$

will be our focus. Combining Lemma 7 with the recovery guarantee of $\mathcal{D}_\ell$, we see that all elements in S_j that survives the subsampling at level ℓ will be recovered. Hence, $\mathbf{Pr}\{i \in R^{(j)}\} \leqslant 2^{-\ell}$ for $i \in S_j^*$ (because it may not be recovered in our range) and $\mathbf{Pr}\{i \in R^{(j)}\} = 2^{-\ell}$ for $i \in S_j^{**}$ (because if it survives the subsampling it would be recovered). Hence

$$\mathbf{E}\, \widetilde{M}'_j = 2^\ell \sum_{i \in S_j^*} f(x_i)\, \mathbf{Pr}\{i \in R^{(j)}\} \leqslant 2^\ell \sum_{i \in S_j^*} f(x_i) 2^{-\ell} = M_j^*$$

and

$$\mathbf{E}\, \widetilde{M}'_j \geqslant 2^\ell \sum_{i \in S_j^{**}} f(x_i)\, \mathbf{Pr}\{i \in R^{(j)}\} = 2^\ell \sum_{i \in S_j^{**}} f(x_i) 2^{-\ell} = M_j^{**}$$

■

Lemma 9 *With probability at least* 0.95 *(over the subsamplings), it holds for all* $j \leqslant j_0$ *that* $(1-O(\varepsilon))M_j^{**} \leqslant \widetilde{M}_j \leqslant (1+O(\varepsilon))M_j^*$.

Proof The argument is similar to the preceding lemma. Note that there are at most $2^{j_0+1} = 4L^3/\varepsilon^3$ elements of interest in this case, and $\mathcal{D}_0$ is guaranteed to recover all of them, since

$$f(x_i) \geqslant \xi 2^{-j_0} f(x) \geqslant \frac{\varepsilon^3}{4L^3} f(x)$$

and we choose $\theta = \min\{\varepsilon^3/(4C_f L^3), \alpha\varepsilon/4, h^{-1}(\alpha\varepsilon/3)\}$ for $\mathcal{D}_0$, where C is an absolute constant. Each $f(x_i)$ is estimated up to an $(1+O(\varepsilon))$-factor. ■

Lemma 10 *With probability at least* 0.8 *(over subsamplings) it holds that*

$$(1-O(\varepsilon))\sum_{j\in\mathcal{J}} M_j^{**} - O(\varepsilon M) \leqslant \sum_{j\in\mathcal{J}} \widetilde{M}_j \leqslant (1+O(\varepsilon))M.$$

Proof Note that $f(x_i)$ are within a factor of 2 from each other for $i \in S_j^*$, thus

$$\sum_{i\in S_j^*} f(x_i)^2 \leqslant \frac{4}{|S_j^*|}\left(\sum_{i\in S_j^*} f(x_i)\right)^2 = \frac{4}{|S_j^*|}(M_j^*)^2.$$

When $j > j_0$ and $j \in \mathcal{J}$, we showed that $|S_j^*| \geqslant \varepsilon^{-2}L^2$ (Lemma 7), thus

$$\sum_{i\in S_j^*} f(x_i)^2 = O\left(\frac{\varepsilon^2}{L^2}(M_j^*)^2\right).$$

It follows from Chebyshev's inequality that with probability at least 0.95,

$$\left|\sum_{j>j_0, j\in\mathcal{J}} \widetilde{M}_j - \mathbf{E}\sum_{j>j_0, j\in\mathcal{J}} \widetilde{M}_j\right| = O\left(\varepsilon \sum_{j>j_0, j\in\mathcal{J}} M_j^*\right)$$

Combining with Lemma 8, we have with probability at least 0.85,

$$(1-O(\varepsilon))\sum_{j>j_0, j\in\mathcal{J}} M_j^{**} - O(\varepsilon)\sum_{j>j_0, j\in\mathcal{J}} M_j^* \leqslant \sum_{j>j_0, j\in\mathcal{J}} \widetilde{M}_j \leqslant (1+O(\varepsilon))\sum_{j>j_0, j\in\mathcal{J}} M_j^*$$

Further combining with Lemma 9, we have with probability at least 0.8,

$$(1-O(\varepsilon))\sum_{j\in\mathcal{J}} M_j^{**} - O(\varepsilon)\sum_{j>j_0, j\in\mathcal{J}} M_j^* \leqslant \sum_{j\in\mathcal{J}} \widetilde{M}_j \leqslant (1+O(\varepsilon))\sum_{j\in\mathcal{J}} M_j^*$$

The result follows from the observation that $\sum_j M_j^* \leqslant M$. ■

Note that the levels $j \notin \mathcal{J}$ contribute at most $O(\varepsilon M)$ in expectation to the total norm. By Markov's inequality, except with probability 0.05 (over subsampling), they contribute at most $O(\varepsilon M)$. Combining with the preceding lemma, we have concluded that with probability at least 0.75,

$$(1 - O(\varepsilon)) \sum_{j \geqslant 1} M_j^{**} - O(\varepsilon M) \leqslant \sum_{j \geqslant 1} \widetilde{M}_j \leqslant (1 + O(\varepsilon))M.$$

Over the randomness of ζ, for each i, with probability at least $1 - O(\varepsilon)$, we have $i \in S_{j'}^{**}$ for some j'. This implies that

$$\mathbf{E}\left(M - \sum_{j \geqslant 1} M_j^{**} \right) = O(\varepsilon M).$$

By Markov's inequality, we have with probability (over ζ) at least 0.95 that

$$(1 - O(\varepsilon))M \leqslant \sum_{j \geqslant 1} M_j^{**} \leqslant M.$$

Finally, combining with the failure probability of the HeavyHitter structures, we conclude that with probability at least 0.7,

$$(1 - O(\varepsilon))M \leqslant \sum_{j \geqslant 1} \widetilde{M}_j \leqslant (1 + O(\varepsilon))M.$$

D.6. Analysis of Algorithm 3 with Bad $\widehat{M}$

We have proved that Algorithm 3, when provided a good overestimate $\widehat{M}$, gives a good estimate $\widetilde{M}$ to M in the preceding Section D.5. In this section, we show that the algorithm does not overestimate when $\widehat{M}$ is bad. We follow the notations in the preceding section and assume likewise that $K = 2$.

Lemma 11 *Suppose that $\widehat{M} > 2M$. Algorithm 3 returns an estimate $\widetilde{M}$, which with probability at least 0.7 (over ζ and subsampling) satisfies that $\widetilde{M} \leqslant (1 + O(\varepsilon))M$.*

Proof There exists $j^* < K$ such that $\widehat{M}^* = \widehat{M}/2^{j^*} \in [M, 2M]$. We compare the behavior of Algorithm 3 on estimate $\widehat{M}$ and $\widehat{M}^*$, under the same randomness in the subsampling functions, heavy hitter structures and ζ. Denote the magnitude levels associated with $\widehat{M}^*$ by $S_1^*, S_2^*, \ldots$ and the levels associated with $\widehat{M}$ by $S_1, S_2 \ldots$. It is clear that $S_1 = \cdots = S_{j^*} = \emptyset$ and $S_j = S_{j-j^*}^*$ for $j > j^*$. Hence for $j \leqslant j_0$, we can still recover all items in $S_1^*, \ldots, S_{j_0^*}^*$ for $j_0^* = j_0 - j^*$, that is, all items in $S_1, \ldots, S_{j_0^*+j^*}$. Observe that $j^* < \log K$ and so $j_0^* + j^* < j_0$, and so it is possible that we miss the levels S_j for $j = j_0^* + j^* + 1, \ldots, j_0$ since the subsequent for-loop starts with S_{j_0+1}. All the recovered levels are within $(1 \pm O(\varepsilon))$-factor of their true values, according to the proof of Theorem 6, with probability at least 0.7. Therefore, we shall never overestimate, that is, $\widetilde{M} \leqslant (1 + O(\varepsilon))M$. ■

Lemma 12 *Suppose that $\widehat{M} < M$. Algorithm 3 returns an estimate $\widetilde{M}$, which with probability at least 0.7 (over ζ and subsampling) satisfies that $\widetilde{M} \leqslant (1 + O(\varepsilon))M$.*

Proof There exists j^* such that $2^{j^*}\widehat{M} = \widehat{M^*} \in [M, 2M]$. Similar to the proof of Lemma 11, we compare the behavior of Algorithm 3 on estimate $\widehat{M}$ and $\widehat{M^*}$, under the same randomness in the subsampling functions, heavy hitter structures and ζ. Let $\{S_j^*\}$ and $\{S_j\}$ be as defined in the proof of Lemma 11. Now we have $S_j = S^*_{j+j^*}$ and may miss the bands $S_1^*, \ldots, S_{j^*}^*$. The rest follows as in Lemma 11. ■

Appendix E. ℓ_1 Subspace Embeddings for i.i.d. Random Design Matrices

In this section we present oblivious ℓ_1 subspace embeddings for i.i.d. random design matrices. This allows us to achieve a polynomial-sized sketch without paying the general case distortion lower bound of $\Omega(d/\log^2 r)$ of Wang and Woodruff (2019).

In consideration of practical applications, we specifically focus on *heavy-tailed distributions*. In fact, as we will see, these are the most interesting from a theoretical perspective as well. Our model for our heavy-tailed distributions will be symmetric power law distributions of index p, which are distributions that satisfy

$$1 - F(x) \sim cx^{-p}$$

for a constant c. In the literature, works such as Zhang and Zhou (2018); Balkema and Embrechts (2018) have considered linear regression in the ℓ_1 norm with heavy tailed i.i.d. design matrices.

p	Distortion upper bound	Distortion lower bound
$p \in (0,1)$	$O(1)$ (Theorem E.8)	1
$p = 1$	$O\left(\frac{\log n}{\log(r/d^2 \log d)}\right)$ (Theorem E.9)	$\Omega\left(\frac{\log n}{\log r}\right)$ (Theorem E.24)
$p \in (1,2)$	$O\left(\frac{d^{1/p}}{(r/d^2)^{1-1/p}}\right)$, $n^{1-1/p} > d^{1/p}\log d$ (Theorem E.18) $O\left(\frac{d^{1/p}\log d}{(r/d^2\log d)^{1-1/p}}\right)$, $n^{1-1/p} \leqslant d^{1/p}\log d$ (Theorem E.19)	$\Omega\left(\frac{d^{1/p}}{r^{1-1/p}}\right)$ (Theorem E.25)
$p \geqslant 2$	$1+\varepsilon$ (Theorem E.23)	1

Table 1: Results for i.i.d. symmetric power law design matrices

Throughout this section, let $\mathcal{D}$ be a symmetric power law distribution with index $p \geqslant 0$, and let $\mathbf{A} \sim \mathcal{D}^{n\times d}$ be a matrix drawn with i.i.d. entries drawn from $\mathcal{D}$, unless noted otherwise.

E.1. Setup for analysis

Let $\mathbf{v} \in \mathbb{R}^n$ be a vector. We will frequently refer to the kth level set $\mathbf{v}_{(k)}$ of $\mathbf{v}$, which takes on the values of $\mathbf{v}$ whenever it has absolute value in $[2^k, 2^{k+1})$, and 0 otherwise.

Definition E.1 (Level sets of a vector) *We define the kth level set $\mathbf{v}_{(k)}$ of $\mathbf{v} \in \mathbb{R}^n$ coordinate-wise by*

$$\mathbf{e}_i^\top \mathbf{v}_{(k)} \coloneqq \begin{cases} \mathbf{e}_i^\top \mathbf{v} & \text{if } |\mathbf{e}_i^\top \mathbf{v}| \in [2^k, 2^{k+1}) \\ 0 & \text{otherwise} \end{cases}.$$

For $k = 0$, we set

$$\mathbf{e}_i^\top \mathbf{v}_{(0)} \coloneqq \begin{cases} \mathbf{e}_i^\top \mathbf{v} & \text{if } |\mathbf{e}_i^\top \mathbf{v}| \in [0, 2) \\ 0 & \text{otherwise} \end{cases}.$$

We will repeatedly make use of the following simple lemmas about CountSketch and symmetric power law distributions.

Lemma E.2 (No expansion) *Let* $\mathbf{S}$ *be drawn as an* $r \times n$ CountSketch *matrix with random signs* $\sigma : [n] \to \{\pm 1\}$ *and hash functions* $h : [n] \to [r]$*. Then for all* $\mathbf{v} \in \mathbb{R}^n$*,*

$$\|\mathbf{S}\mathbf{v}\|_1 \leqslant \|\mathbf{v}\|_1.$$

Proof

$$\|\mathbf{S}\mathbf{v}\|_1 = \sum_{i=1}^{r} \left| \sum_{j:h(j)=i}^{n} \sigma_i v_i \right| \leqslant \sum_{i=1}^{r} \sum_{j:h(j)=i}^{n} |\sigma_i v_i| = \sum_{j=1}^{n} |v_i| = \|\mathbf{v}\|_1$$

■

Lemma E.3 *Let* $\mathcal{D}$ *be a symmetric power law distribution with index* $p > 0$*. Then, for* k *a large enough constant depending on* $\mathcal{D}$*,*

$$\mathop{\mathbf{Pr}}_{X \sim \mathcal{D}}\Big(|X| \in [2^k, 2^{k+1})\Big) = \Theta(2^{-kp})$$

and

$$\mathop{\mathbf{Pr}}_{\mathbf{v} \sim \mathcal{D}^n}\Big(\|\mathbf{v}_{(k)}\|_0 = \Theta(n2^{-kp})\Big) \geqslant 1 - \exp\Big(\Theta(n2^{-kp})\Big)$$

Proof For a large enough k, we have that

$$\mathop{\mathbf{Pr}}_{X \sim \mathcal{D}}\Big(|X| \in [2^k, 2^{k+1})\Big) = \overline{F}(2^k) - \overline{F}(2^k) = \Theta\left(\frac{1}{2^{kp}} - \frac{1}{2^{(k+1)p}}\right) = \Theta(2^{-kp}).$$

Then in expectation,

$$\|\mathbf{v}_{(k)}\|_0 = \Theta(n2^{-kp})$$

so we conclude by Chernoff bounds. ■

Lemma E.4 *Let* $\mathcal{D}$ *be a symmetric power law distribution with index* $p > 0$*. Then,*

$$\mathop{\mathbf{Pr}}_{\mathbf{A} \sim \mathcal{D}^{n \times d}}\Big(\|\mathbf{A}\|_\infty \leqslant O\Big((nd/\delta)^{1/p}\Big)\Big) \geqslant 1 - \delta$$

Proof Each entry is at most $O((nd/\delta)^{1/p})$ with probability at least δ/nd, so we conclude by a union bound over the nd entries. ■

Definition E.5 (Truncation) *For* $T > 0$ *and* $x \in \mathbb{R}$*, define*

$$\mathsf{trunc}_T(x) \coloneqq \begin{cases} x & |x| \leqslant T \\ 0 & \textit{otherwise} \end{cases}.$$

For a distribution $\mathcal{D}$*, we define* $\mathsf{trunc}_T(\mathcal{D})$ *to be the distribution that draws* $\mathsf{trunc}_T(X)$ *for* $X \sim \mathcal{D}$*.*

Lemma E.6 (Moments of truncated power laws) *Let $\mathcal{D}$ be a power law distribution with index $p > 0$. Let $T > 0$ be sufficiently large. Then*

$$\mathop{\mathbf{E}}_{X \sim \mathsf{trunc}_T(\mathcal{D})} |X| = \begin{cases} \Theta(T^{1-p}) & \text{if } p \in (0,1) \\ \Theta(\log T) & \text{if } p = 1 \\ \Theta(1) & \text{if } p > 1 \end{cases}$$

$$\mathop{\mathbf{E}}_{X \sim \mathsf{trunc}_T(\mathcal{D})} X^2 = \begin{cases} \Theta(T^{2-p}) & \text{if } p \in (0,2) \\ \Theta(\log T) & \text{if } p = 2 \\ \Theta(1) & \text{if } p > 2 \end{cases}$$

Proof The proof is deferred to Appendix I. ■

Definition E.7 *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $T > 0$. Then, we write $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ where $\mathbf{A}^H$ is the submatrix of $\mathbf{A}$ formed by the rows containing an entry with absolute value at least T, and $\mathbf{A}^L$ is the rest of the rows.*

E.2. Algorithms for $p < 1$

We first present the results that for tails that are very heavy admit $O(1)$ distortion embeddings in $\mathrm{poly}(d)$ dimensions for a very simple reason: when $p < 1$, then the largest entry in every vector is a good approximation of the entire ℓ_1 mass of the vector.

Theorem E.8 *Let $\mathcal{D}$ be a symmetric power law distribution with index $p \in (0,1)$. Let $\mathbf{S}$ be drawn as a CountSketch matrix with $r = O(d^2 \log^2 d)$ rows. Then*

$$\Pr_{\mathbf{S},\mathbf{A}}\left(\Omega(1)\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{A}\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\right) \geqslant \frac{99}{100}.$$

Proof The proof proceeds similarly to the case of $p = 1$, and is deferred to Appendix I. ■

Thus, we focus on the regime of $p \geqslant 1$.

E.3. Algorithms for $p = 1$

In this section, we prove the following:

Theorem E.9 *Let $\mathcal{D}$ be a symmetric power law distribution with index $p = 1$. Let $\mathbf{S} \in \mathbb{R}^{r \times n}$ be drawn as a CountSketch matrix. Then, for any $C(d \log d)^2 \leqslant r \leqslant o(\sqrt{n})$ for C a large enough constant, we have*

$$\Pr_{\mathbf{A} \sim \mathcal{D}^{n \times d}}\left\{\frac{1}{\kappa}\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{A}\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\right\} \geqslant \frac{99}{100}$$

for

$$\kappa = O\left(\frac{\log n}{\log(r/d^2 \log d)}\right).$$

The idea is that with r rows of CountSketch, we can preserve the top r entries of $\mathbf{A}\mathbf{x}$, which has mass approximately $\Omega(n \log r)\|\mathbf{x}\|_1$, while the rest of the entries have mass at most $O(n \log n)\|\mathbf{x}\|_1$. We formalize this idea in the following several lemmas.

Lemma E.10 (Mass of small entries) *Let $\mathcal{D}$ be a power law distribution with index $p = 1$ and let $\mathbf{A} \sim \mathcal{D}^{n \times d}$. Let $\mathrm{poly}(d) \leqslant T \leqslant n$ and let $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ as in Definition E.7. Then,*

$$\mathbf{Pr}\Big(\|\mathbf{A}^L\mathbf{x}\|_1 \leqslant O(n \log T)\|\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\Big) \geqslant 0.99$$

Proof Note that $\mathbf{A}^L$ is drawn i.i.d. from $\mathsf{trunc}_T(\mathcal{D})$ so by Lemma E.6, it has entries with first two moments

$$\mu \coloneqq \mathop{\mathbf{E}}_{X \sim \mathcal{D}}(|X| \mid |X| \leqslant T) = \Theta(\log T)$$
$$\sigma^2 \coloneqq \mathop{\mathbf{E}}_{X \sim \mathcal{D}}\big(X^2 \mid |X| \leqslant T\big) = \Theta(T)$$

Then for a single column $\mathbf{A}^L\mathbf{e}_j$ for $j \in [d]$, by Bernstein's inequality,

$$\mathbf{Pr}\big(\|\mathbf{A}^L\mathbf{e}_j\|_1 \geqslant 2\mu n\big) \leqslant \exp\left(-\frac{1}{2}\frac{(\mu n)^2}{\sigma^2 n + \mu n T/3}\right) \leqslant \exp(-\Omega(\log T)) = \frac{1}{\mathrm{poly}(T)}.$$

Then since $T \geqslant \mathrm{poly}(d)$, we may union bound over the d columns so that $\|\mathbf{A}^L\mathbf{e}_j\|_1 = O(\mu n) = O(n \log T)$ for all columns $j \in [d]$ with probability at least $1 - 1/\mathrm{poly}(d)$. Conditioned on this event, we have for all $\mathbf{x} \in \mathbb{R}^d$ that

$$\|\mathbf{A}^L\mathbf{x}\|_1 \leqslant \sum_{j=1}^{d} |\mathbf{x}_j| \|\mathbf{A}^L\mathbf{e}_j\|_1 = O(n \log T)\|\mathbf{x}\|_1.$$

■

Lemma E.11 (Unique hashing of large entry rows) *Let $\mathcal{D}$ be a power law distribution with index $p > 0$ and cdf F, and let $\mathbf{A} \sim \mathcal{D}^{n \times d}$. Let t and r be parameters such that $r \geqslant C(d \log d)^2$ for a sufficiently large constant C and $r = o(\sqrt{n})$, and define*

$$\tau_1 \coloneqq F^{-1}\left(1 - \frac{C' \log d}{n}\right) = \Theta((n/\log d)^{1/p})$$
$$\tau_2 \coloneqq F^{-1}\left(1 - \frac{r/d^2 \log d}{n}\right) = \Theta((nd^2 \log d/r)^{1/p})$$
$$\mathcal{R}_1 \coloneqq \Big\{i \in [n] : \exists j \in [d], |\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j| > \tau_1\Big\}$$
$$\mathcal{R}_2 \coloneqq \Big\{i \in [n] : \exists j \in [d], |\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j| > \tau_2\Big\}$$

for a sufficiently large constant C'. Then if we hash each row of $\mathbf{A}$ into $O(r)$ hash buckets, with probability at least 0.95:

– *Every row of $\mathcal{R}_1$ is hashed to a bucket with no other row from $\mathcal{R}_2$.*

– *For every column $j \in [d]$ and every integer $\log_2 \tau_2 \leqslant k \leqslant \log_2 \tau_1$ has $\Theta(n/2^{kp})$ rows with a large entry with absolute value in $[2^k, 2^{k+1})$ that are hashed to a bucket with no other row from $\mathcal{R}_2$.*

– *Let $\mathcal{R}$ be the set of rows which are hashed with no other row from $\mathcal{R}_2$, which we refer to as* uniquely hashed *rows. Then $|\mathcal{R}| = \Theta(r/d \log d)$ and every one of these large entries is on a distinct row.*

Proof For each $j \in [d]$, the number of expected entries in the jth column with absolute value at least τ_1 is $C' \log d$, so by Chernoff bounds, with probability at least $1 - \exp(\Theta(\log d)) = 1 - 1/\operatorname{poly}(d)$, there are at most $O(\log d)$ such entries. By a union bound over the d columns, this is true for all d columns with probability at least 0.99.

Similarly, for $\log_2 \tau_2 \leqslant k \leqslant \log_2 \tau_1$, we have by Lemma E.3 that

$$\mathbf{Pr}\Big(\|(\mathbf{A}\mathbf{e}_j)_{(k)}\|_0 = \Theta(n/2^{kp})\Big) \geqslant 1 - \exp\Big(\Theta(n/2^{kp})\Big).$$

Then summing over k, we have that

$$\begin{aligned}\mathbf{Pr}\left(\bigcap_{k=\log_2 \tau_2}^{\log_2 \tau_1} \Big\{\|(\mathbf{A}\mathbf{e}_j)_{(k)}\|_0 = \Theta(n/2^{kp})\Big\}\right) &\geqslant 1 - \sum_{k=\log_2 \tau_2}^{\log_2 \tau_1} \exp(\Theta(n/2^{kp})) \\ &\geqslant 1 - \exp(\Theta(C' \log d)) = 1 - \frac{1}{\operatorname{poly}(d)}.\end{aligned}$$

By a union bound over the d columns, every large entry level set of every column $\mathbf{A}\mathbf{e}_j$ has the expected number of elements, up to constant factors, simultaneously with probability at least $99/100$. Conditioned on this event, $|\mathcal{R}_1| = O(d \log d)$.

Note that across the d columns, there are $\Theta(r/d \log d)$ rows corresponding to level sets k for $\log_2 \tau_2 \leqslant k \leqslant \log_2 \tau_1$. Then with $O(r)$ hash buckets, each pair of rows from $\mathcal{R}_1 \times \mathcal{R}_2$ is hashed to a separate bucket with probability $O(1/|\mathcal{R}_1 \times \mathcal{R}_2|) = O(1/r)$, so every row in $\mathcal{R}_1$ is uniquely hashed with probability 0.99 by a union bound. Furthermore, with $O(r) = \omega(r/d \log d)$ hash buckets, we have by Lemma A.4 that for each level set $\mathbf{v}_{(k)}$, half of the $\Theta(n/2^{kp})$ rows in the kth level set get hashed to a bucket with no other row from $\mathcal{R}$ with probability at least

$$1 - 2\exp\left(\frac{(1/2)^2}{12}\Theta(n/2^{kp})\right) = 1 - \exp\Big(\Theta(n/2^{kp})\Big).$$

Then again by a union bound over the level sets and columns, every large entry level set of every column has at least half of their rows hashed with no other row from $\mathcal{R}$, simultaneously with probability at least $99/100$.

The probability that any two of the large entries lie on the same row is $O(|\mathcal{R}_2|^2/n) = o(1)$. Then by a union bound, the total success probability for the entire lemma is at least 0.95. ∎

We apply the lemma above to show that when we write $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ as in Definition E.7, then $\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 = \Omega(n \log(r/d))\|\mathbf{x}\|_1$ for all $\mathbf{x} \in \mathbb{R}^d$ when we choose $T = nd^2 \log d/r$.

Lemma E.12 (Mass of large entries) *Let Let* $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ *as in Definition E.7 with* $T = nd^2 \log d/r$. *Let* $\mathbf{S}$ *be a* CountSketch *matrix with* r *rows. Then with probability at least* 0.95,

$$\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 \geqslant \|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 \geqslant \Omega(\|\mathbf{A}^H\mathbf{x}\|_1) \geqslant \Omega(n\log(r/d^2\log d))\|\mathbf{x}\|_1$$

for all $\mathbf{x} \in \mathbb{R}^d$, *where* $\mathbf{B}'$ *is the subset of uniquely hashed rows of* $\mathbf{A}^H$ *given by Lemma E.11.*

Proof Let $\mathbf{B}'$ be the subset of rows of $\mathbf{A}^H$ given by Lemma E.11 that are hashed to locations without any other rows of $\mathbf{A}^H$. Recall also τ_1 and τ_2 from the lemma.

We first have that $\|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 = \Omega(\|\mathbf{A}^H\mathbf{x}\|_1)$ since the rows containing entries larger than τ_1 are perfectly hashed, while rows containing entries between τ_2 and τ_1 are preserved up to constant factors.

Let $\mathbf{B}' = \mathbf{B}'_{>T} + \mathbf{B}'_{\leqslant T}$ where $\mathbf{B}'_{>T}$ contains the entries of $\mathbf{B}'$ that have absolute value greater than T and $\mathbf{B}'_{\leqslant T}$ contains the rest of the entries. Note then that $\mathbf{B}'_{>T}$ has at most one nonzero entry per row, and $\mathbf{B}'_{\leqslant T}$ has at most $O(d \cdot r/d\log d) = O(r/\log d)$ nonzero entries and thus by Lemma E.4, $\|\mathbf{B}'_{\leqslant T}\|_\infty \leqslant O(r)$ with probability at least 0.99. We condition on this event. Then for all $\mathbf{x}$,

$$\begin{aligned}
\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 &\geqslant \|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 \\
&\geqslant \|\mathbf{S}\mathbf{B}'_{>T}\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 \\
&= \sum_{j=1}^{d} |\mathbf{x}_j| \|\mathbf{B}'_{>T}\mathbf{e}_j\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \text{Since the } \mathbf{B}'_{>T}\mathbf{e}_j \text{ have disjoint support} \\
&\geqslant \sum_{j=1}^{d} |\mathbf{x}_j| \sum_{k=\log_2\tau_2}^{\log_2\tau_1} 2^k \Theta(n/2^k) - \|\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \text{Lemmas E.11 and E.2} \\
&= \Omega(n(\log_2\tau_1 - \log_2\tau_2))\|\mathbf{x}\|_1 - O(r)\|\mathbf{B}'_{\leqslant T}\|_\infty\|\mathbf{x}\|_1 && \text{Hölder's inequality} \\
&= \Omega(n\log(r/d^2\log d))\|\mathbf{x}\|_1 - O(r^2)\|\mathbf{x}\|_1 \\
&= \Omega(n\log(r/d^2\log d))\|\mathbf{x}\|_1
\end{aligned}$$

as desired. ■

The last thing we need to bound is the mass contribution of the rows of $\mathbf{A}^L$ that are hashed together with the uniquely hashed rows of $\mathbf{A}^H$. We first bound the columns of the matrix $\mathbf{S}'\mathbf{A}^L$, where $\mathbf{S}'$ is a subset of hash buckets.

Lemma E.13 *Let* $\mathcal{D}$ *be a symmetric power law distribution with index* $p \in (0,2)$ *with cdf* F. *Let* $T := F^{-1}(1 - r/nd^2\log d) = \Theta((nd^2\log d/r)^{1/p})$, $r' < r$. *Let* $\mathbf{S}'$ *be a subset of* r' *rows of a* $r \times n$ *CountSketch matrix* $\mathbf{S}$. *Let* $\mathbf{C} \sim \mathsf{trunc}_T(\mathcal{D})^{n\times d}$. *Then for each* $j \in [d]$,

$$\mathbf{Pr}\Big(\|\mathbf{S}'\mathbf{C}\mathbf{e}_j\|_1 \leqslant O\Big(r' + \lambda\sqrt{r'}\Big)(d^2\log d)^{1/p-1/2}(n/r)^{1/p}\Big) \geqslant 1 - \frac{1}{\lambda}.$$

Proof The proof is just a second moment bound and is deferred to Appendix I. ■

Given the above bounds, the rest of the proof is just Hölder's inequality.

Lemma E.14 *Let $r \geqslant C(d\log d)^2$ for a large enough constant C. Let $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ as in Definition E.7 with $T = (nd^2\log d/r)^{1/p}$. Let $\mathbf{S}$ be a CountSketch matrix with r rows. Let $\mathbf{A}^L = \mathbf{C}_1 + \mathbf{C}_2$, where $\mathbf{C}_1$ is the submatrix formed by the rows of $\mathbf{A}^L$ that are hashed together with the uniquely hashed rows of $\mathbf{A}^H$ by $\mathbf{S}$ (c.f. Lemma E.11), and $\mathbf{C}_2$ is the submatrix formed by the rest of the rows. Then with probability at least 0.99, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\|\mathbf{S}\mathbf{C}_1\mathbf{x}\|_1 \leqslant O\left(\frac{1}{\sqrt{\log d}}\frac{n^{1/p}}{(r/d^2\log d)^{1/p-1}}\right)\|\mathbf{x}\|_1.$$

Proof Let $\mathbf{S}'$ be the submatrix of $\mathbf{S}$ formed by the set of r' uniquely hashed rows from Lemma E.11, with $r' = O(r/d\log d)$. Setting $\lambda = 100d$, we have that

$$r' + \lambda\sqrt{r'} = O(r') = O\left(\frac{r}{d\log d}\right)$$

so

$$\mathbf{Pr}\left(\|\mathbf{S}\mathbf{C}_1\mathbf{e}_j\|_1 \leqslant O(r')(d^2\log d)^{1/p-1/2}\left(\frac{n}{r}\right)^{1/p}\right) = \mathbf{Pr}\left(\|\mathbf{S}'\mathbf{A}^L\mathbf{e}_j\|_1 \leqslant O\left(\frac{1}{\sqrt{\log d}}\frac{n^{1/p}}{(r/d^2\log d)^{1/p-1}}\right)\right)$$
$$\geqslant 1 - \frac{1}{100d}.$$

By a union bound over the d columns, this is true for all $k \in [d]$ with probability at least 0.99. Conditioned on this event, we have by the triangle inequality that

$$\|\mathbf{S}\mathbf{C}_1\mathbf{x}\|_1 \leqslant \sum_{k=1}^{d}|\mathbf{x}_k|\sum_{i\in S}Y_{i,k} \leqslant O\left(\frac{1}{\sqrt{\log d}}\frac{n^{1/p}}{(r/d^2\log d)^{1/p-1}}\right)\|\mathbf{x}\|_1$$

as desired. ∎

With the above lemmas in place, we prove Theorem E.9.

Proof [Proof of Theorem E.9] The "no dilation" bound is just Lemma E.2. We thus focus on the "no contraction" bound.

We condition on the results of Lemmas E.10, E.12, and E.14. By a union bound, these all hold simultaneously with probability at least 0.9. Then, we have for all $\mathbf{x}$ that

$$\frac{1}{\kappa} \geqslant \frac{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{A}\mathbf{x}\|_1} \geqslant \frac{\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{C}_1\mathbf{x}\|_1}{\|\mathbf{A}^H\mathbf{x}\|_1 + \|\mathbf{A}^L\mathbf{x}\|_1} \geqslant \frac{\Omega(\|\mathbf{A}^H\mathbf{x}\|_1 + n\log(r/d^2\log d))\|\mathbf{x}\|_1}{O(\|\mathbf{A}^H\mathbf{x}\|_1 + n\log(nd^2\log d/r))\|\mathbf{x}\|_1} \geqslant \Omega\left(\frac{\log(r/d^2\log d)}{\log n}\right).$$

∎

E.4. Algorithms for $p \in (1,2)$

For power law distributions with index $p \in (1,2)$, we need different algorithms based on the parameter regime: when n is rather large, then the distribution looks relatively flat so that sampling is approximately optimal, while when n is rather small, then the variance is large enough so that CountSketch helps capture and preserve large values that make up a significant fraction of the mass.

E.4.1. LARGE n: SAMPLING

When n is large, we shall see that by concentration, sampling alone will give us nearly tight distortion bounds.

We first prove concentration in the upper tail.

Lemma E.15 (Upper tail concentration) *Let $\mathcal{D}$ be a power law distribution with index $p \in (1,2)$ and let $\mathbf{A} \sim \mathcal{D}^{n\times d}$. Then,*

$$\mathbf{Pr}\left\{\|\mathbf{Ax}\|_1 \leqslant O\left(1 + \frac{d^{1/p}\log d}{n^{1-1/p}}\right)\|\mathbf{x}\|_1 n, \forall \mathbf{x} \in \mathbb{R}^d\right\} \geqslant 0.99.$$

Proof By a union bound, $\|\mathbf{A}\|_\infty \leqslant B = O((nd)^{1/p})$ with probability at least 0.999. Conditioned on this event, each entry of $\mathbf{A}$ is distributed as $\mathsf{trunc}_B(\mathcal{D})$. Note that for a random variable $X \sim \mathsf{trunc}_B(\mathcal{D})$, we have by Lemma E.6 that

$$\mathbf{E}|X| = \Theta(1)$$
$$\mathbf{E}|X|^2 = \Theta(B^{2-p})$$

Now let $\mathbf{v} \sim \mathsf{trunc}_B(\mathcal{D})^n$. By the upper tail Bernstein bound,

$$\mathbf{Pr}\{\|\mathbf{v}\|_1 - n\,\mathbf{E}|X| \geqslant \lambda\} \leqslant \exp\left(-\frac{1}{2}\frac{\lambda^2}{B^{2-p}n + B\lambda/3}\right).$$

Then with $\lambda = \kappa n$ for

$$\kappa = \frac{d^{1/p}\log d}{n^{1-1/p}},$$

we have

$$\lambda = \frac{d^{1/p}\log d}{n^{1-1/p}}n = (nd)^{1/p}\log d \geqslant B\log d$$

and

$$\lambda^2 = \frac{d^{2/p}\log^2 d}{n^{2-2/p}}n^2 = (nd)^{2/p}\log^2 d = (nd)^{\frac{1}{p}(2-p)}(nd)\log^2 d \geqslant B^{2-p}n\log d$$

so we have that with probability at least $1 - 1/\operatorname{poly}(d)$,

$$\|\mathbf{v}\|_1 - n\,\mathbf{E}|X| \leqslant \lambda \implies \|\mathbf{v}\|_1 \leqslant n\,\mathbf{E}|X| + \kappa n = \Theta\left(1 + \frac{d^{1/p}\log d}{n^{1-1/p}}\right)n.$$

By a union bound over the d columns of $\mathbf{A}$, this holds simultaneously for all columns of $\mathbf{A}$ with probability at least $1 - 1/\operatorname{poly}(d)$. We condition on this event. It then follows that for every $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{Ax}\| \leqslant \|\mathbf{x}\|_1 \max_{j=1}^{d}\|\mathbf{A}\mathbf{e}_j\|_1 \leqslant \Theta\left(1 + \frac{d^{1/p}\log d}{n^{1-1/p}}\right)n.$$

■

To prove concentration in the lower tail, we first need the following lemma.

Lemma E.16 *Let $\mathcal{D}$ be a symmetric power law distribution with index $p \in (1,2)$. Let $\mathbf{x} \in \mathbb{R}^d$. Then*

$$\Pr_{\mathbf{v}\sim\mathcal{D}^d}(|\langle \mathbf{v}, \mathbf{x}\rangle| \geqslant \Omega(\|\mathbf{x}\|_p)) = \Omega(1). \tag{4}$$

Proof The proof is distracting from this discussion and is deferred to Appendix I. ■

Lemma E.17 (Lower tail concentration) *Let $\mathcal{D}$ be a symmetric power law distribution with index $p \in (1,2)$. Let $n \geqslant d\log d$. Then,*

$$\Pr_{\mathbf{A}\sim\mathcal{D}^{n\times d}}\left\{\|\mathbf{A}\mathbf{x}\|_1 \geqslant \Omega(n\|\mathbf{x}\|_p), \forall \mathbf{x} \in \mathbb{R}^d\right\} \geqslant 0.99.$$

Proof Let $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_1 = 1$ and let $\mathbf{v} \sim \mathcal{D}^d$. Then by Lemma E.16,

$$\Pr_{\mathbf{v}\sim\mathcal{D}^d}(|\langle \mathbf{v}, \mathbf{x}\rangle| \geqslant \Omega(\|\mathbf{x}\|_p)) = \Omega(1).$$

Then by Chernoff bounds, at least $\Omega(n)$ of the n rows of $\mathbf{A}\mathbf{x}$ are at least $\Omega(\|\mathbf{x}\|_p)$ with probability at least $1 - \exp(n) = 1 - \exp(d\log d)$. We conclude by a standard net argument. ■

We put the above two parts together for a sketching algorithm based on sampling.

Theorem E.18 *Let $n \geqslant d\log d$ and let $\mathbf{A}$ be drawn as an $n \times d$ matrix of i.i.d. draws from a p-stable distribution. Let*

$$\kappa_n = \Theta\left(1 + \frac{d^{1/p}\log d}{n^{1-1/p}}\right)$$
$$\kappa_r = \Theta\left(1 + \frac{d^{1/p}\log d}{r^{1-1/p}}\right)$$

be the distortion upper bound from Lemma E.15 for when the number of rows is n and r, respectively. Let $\mathbf{S} \in \mathbb{R}^{r\times n}$ be the matrix that samples r rows of $\mathbf{A}$, and then scales by $\kappa_n d^{1-1/p} n/r$. Then,

$$\Pr\left\{\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \kappa_n\kappa_r d^{2(1-1/p)}\|\mathbf{A}\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\right\} \geqslant 0.9$$

In particular, if

$$n^{1-1/p} \geqslant d^{1/p}\log d \iff n \geqslant d^{\frac{1}{p-1}}\log^{\frac{p}{p-1}} d,$$

then

$$\Pr\left\{\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant O\left(1 + \frac{d^{1/p}}{(r/d^2)^{1-1/p}}\right), \forall \mathbf{x} \in \mathbb{R}^d\right\} \geqslant 0.9.$$

Proof By applying lemmas E.15 and E.17, we have that for all $\mathbf{x}$,

$$\Omega(1)\|\mathbf{x}\|_p n \leqslant \|\mathbf{A}\mathbf{x}\|_1 \leqslant \kappa_n\|\mathbf{x}\|_1 n.$$

Furthermore, we can apply the lemmas to $\mathbf{S}\mathbf{A}$ as well, which gives us

$$\Omega(1)\kappa_n d^{1-1/p}\|\mathbf{x}\|_p n \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \kappa_r\kappa_n d^{1-1/p}\|\mathbf{x}\|_1 n.$$

By Hölder's inequality, we have that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{x}\|_p \leqslant \|\mathbf{x}\|_1 \leqslant d^{1-1/p}\|\mathbf{x}\|_p.$$

Thus,

$$\|\mathbf{A}\mathbf{x}\|_1 \leqslant \kappa_n\|\mathbf{x}\|_1 n \leqslant \kappa_n d^{1-1/p}\|\mathbf{x}\|_p n \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1$$

so the sketch does not underestimate norms. On the other hand,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \kappa_r\kappa_n d^{1-1/p}\|\mathbf{x}\|_1 n \leqslant \kappa_r\kappa_n d^{2(1-1/p)}\|\mathbf{x}\|_p n \leqslant \kappa_r\kappa_n d^{2(1-1/p)}\|\mathbf{A}\mathbf{x}\|_1$$

so the sketch does not overestimate norms by more than $\kappa_r\kappa_n d^{2(1-1/p)}$, as claimed. ■

E.4.2. SMALL n: CountSketch

In the previous section, we have handled the case when $n^{1-1/p} \geqslant d^{1/p}\log d$. On the other hand, when $n^{1-1/p} \leqslant d^{1/p}\log d$ we will instead use CountSketch to hash the largest entries of each column of $\mathbf{A}$. These entries are of size around $n^{1/p}$, while the entries of vectors with size smaller than this have mass at most n. Thus, we approximate the mass up to a factor of

$$\frac{n}{n^{1/p}} = n^{1-1/p} \leqslant d^{1/p}\log d,$$

which is roughly what we are shooting for.

Theorem E.19 *Let $n^{1-1/p} \leqslant d^{1/p}\log d$. Let $\mathbf{S}$ be drawn as a CountSketch matrix with r rows. Then*

$$\Pr_{\mathbf{S},\mathbf{A}}\left(\Omega\left(\frac{1}{\kappa}\right)\|\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leqslant \|\mathbf{A}\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\right) \geqslant \frac{99}{100}.$$

where

$$\kappa = O\left(\left(\frac{n}{(r/d^2\log d)}\right)^{1-1/p}\right) = O\left(\frac{d^{1/p}\log d}{(r/d^2\log d)^{1-1/p}}\right).$$

Proof The distortion upper bound again is just Lemma E.2. The distortion lower bound argument is similar to the one presented for the cases of $p \in (0, 1]$ and thus is deferred to Appendix I. ■

E.5. Algorithms for $p \geqslant 2$

When $p \geqslant 2$, we show that any $m \times d$ i.i.d. matrix with $m \geqslant \text{poly}(d)$ has with constant probability, $\|\mathbf{A}\mathbf{x}\|_1 = \Theta(m\|\mathbf{x}_2\|)$ for all $\mathbf{x} \in \mathbb{R}^d$. This shows that a uniform sampling matrix with $\text{poly}(d)$ rows works as a sketch. The following result shows this in expectation.

Lemma E.20 *Let $p \geqslant 2$ and let $\mathcal{D}$ be a symmetric power law with index p. Let $\mathbf{x} \in \mathbb{R}^d$. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}^d}|\langle \mathbf{v}, \mathbf{x}\rangle| = \Theta(\|\mathbf{x}\|_2)$$

Proof The proof is standard and is deferred to Appendix I. ■

Our strategy then is to show that conditioned on every entry of $\mathbf{v} \sim \mathcal{D}^d$ being smaller than some large value $B \geqslant \mathrm{poly}(d)$, the expectation remains approximately unchanged. We then use this in a Bernstein bound to argue the result with high enough probability to union bound over a net.

Lemma E.21 *Let $p \geqslant 2$ and let $\mathcal{D}$ be a symmetric power law with index p. Let $\varepsilon \in (0, 1/4)$ and let $B > \max\{\varepsilon^{-1}, (d\sqrt{d}/\varepsilon)^{1/p}\}$. Let $\mathbf{v} \sim \mathcal{D}^d$ and fix a vector $\mathbf{x} \in \mathbb{R}^d$. Define the events*

$$\mathcal{E}_i \coloneqq \{|\mathbf{v}_i| \leqslant B\}$$
$$\mathcal{E} \coloneqq \bigcap_{i=1}^{d} \mathcal{E}_i$$

Then for B large enough,

$$\mathbf{E}(|\langle \mathbf{v}, \mathbf{x}\rangle| \mid \mathcal{E}) \geqslant (1 - O(\varepsilon))\, \mathbf{E}|\langle \mathbf{v}, \mathbf{x}\rangle|.$$

Proof The proof is standard and is deferred to Appendix I. ■

The following lemma implements the Bernstein bound and applies a standard net argument.

Lemma E.22 *Let $p \geqslant 2$ and let $\mathcal{D}$ be a symmetric power law with index p. Let $\mathbf{A} \sim \mathcal{D}^{m \times d}$ with $m \geqslant \Theta(\max\{\varepsilon^{-p}, (d^{3/2+1/p}\varepsilon^{-1}\log\varepsilon^{-1})^{p/(p-1)}\})$. For $\mathbf{x} \in \mathbb{R}^d$, let $\mu_{\mathbf{x}} \coloneqq \mathbf{E}_{\mathbf{v}\sim\mathcal{D}^d}\langle \mathbf{v}, \mathbf{x}\rangle$. Then,*

$$\mathbf{Pr}\Big(\|\mathbf{A}\mathbf{x}\|_1 = (1 \pm O(\varepsilon))m\mu_{\mathbf{x}}, \forall \mathbf{x} \in \mathbb{R}^d\Big) \geqslant 0.95.$$

Proof Note that with probability at least 0.99, $\|\mathbf{A}\|_\infty = O((md)^{1/p})$. Let this event be $\mathcal{E}$. Then conditioned on $\mathcal{E}$, $\mathbf{A}$ is distributed as an i.i.d. matrix drawn from $\mathcal{D}'$, where $\mathcal{D}'$ is the truncation of $\mathcal{D}$ at

$$B = O((md)^{1/p}) \geqslant \max\Big\{\varepsilon^{-1}, (d\sqrt{d}/\varepsilon)^{1/p}\Big\}$$

where the bound on B follows by our choice of m.

High probability bounds. Now fix $\mathbf{x} \in \mathbb{R}^d$. By Lemmas E.20 and E.21,

$$\mu \coloneqq \mathbf{E}(\|\mathbf{A}\mathbf{x}\|_1 \mid \mathcal{E}) = \sum_{i=1}^{m} \mathbf{E}\Big(|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}| \mid \mathcal{E}\Big) = (1 \pm \varepsilon)\sum_{i=1}^{m} \mathbf{E}|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}| = (1 \pm \varepsilon)m\mu_{\mathbf{x}} = \Theta(m\|\mathbf{x}\|_2) \tag{5}$$

and

$$\sigma^2 \coloneqq \mathbf{Var}(\|\mathbf{A}\mathbf{x}\|_1 \mid \mathcal{E}) = \sum_{i=1}^{m} \mathbf{Var}(|\mathbf{e}_i^\top \mathbf{A}\mathbf{x}| \mid \mathcal{E}) \leqslant O(m\|\mathbf{x}\|_2^2).$$

Then by Bernstein bounds, we have that

$$\mathbf{Pr}(|\|\mathbf{A}\mathbf{x}\|_1 - \mu| \geqslant \varepsilon\mu \mid \mathcal{E}) \leqslant 2\exp\left(-\frac{1}{2}\frac{(\varepsilon\mu)^2}{\sigma^2 + \varepsilon\mu B/3}\right) \leqslant \exp\big(-\Theta(d\log\varepsilon^{-1})\big)$$

where the last inequality follows by our choice of m. Then chaining together with Equation 5,

$$\mathbf{Pr}(\|\mathbf{A}\mathbf{x}\|_1 = (1 \pm O(\varepsilon))m\mu_{\mathbf{x}} \mid \mathcal{E}) \leqslant \exp\big(-\Theta(d\log\varepsilon^{-1})\big). \tag{6}$$

Net argument. We now proceed by a standard net argument. Recall the set $\mathcal{S}$ and the ε-net $\mathcal{N}$ as given in Lemma A.3. Now conditioned on $\mathcal{E}$, we have by a union bound that $\|\mathbf{Ax}\|_1 = (1 \pm O(\varepsilon))m\mu_{\mathbf{x}}$ for every $\mathbf{Ax} \in \mathcal{N}$, with probability at least 0.99. We condition on this event as well. Now for any $\mathbf{y} \in \mathcal{S}$, write $\mathbf{y} = \sum_{i=0}^{\infty} \mathbf{y}^{(i)}$ as given in Lemma A.3. Then,

$$\|\mathbf{SAx}\|_1 = \left\|\mathbf{S}\sum_{i=0}^{\infty}\mathbf{y}^{(i)}\right\|_1 \leqslant \sum_{i=0}^{\infty}\left\|\mathbf{Sy}^{(i)}\right\|_1 \leqslant (1+O(\varepsilon))\sum_{i=0}^{\infty}\left\|\mathbf{y}^{(i)}\right\|_1 \leqslant (1+O(\varepsilon))\sum_{i=0}^{\infty}\varepsilon^i \leqslant 1+O(\varepsilon).$$

We conclude by homogeneity. ■

Given the above lemma, our subspace embedding follows simply from uniform sampling and rescaling.

Theorem E.23 *Let $p \geqslant 2$ and let $\mathcal{D}$ be a symmetric power law with index p. Let $\varepsilon \in (0, 1/2)$, let*

$$r = \Theta(\max\left\{\varepsilon^{-p}, (d^{3/2+1/p}\varepsilon^{-1}\log\varepsilon^{-1})^{p/(p-1)}\right\})$$

and let $\mathbf{S} \in \mathbb{R}^{r\times n}$ be a matrix that uniform samples r rows and scales by n/r. Then,

$$\mathbf{Pr}\Big(\|\mathbf{SAx}\|_1 = (1 \pm O(\varepsilon))\|\mathbf{Ax}\|_1, \forall \mathbf{x} \in \mathbb{R}^d\Big) \geqslant 0.9$$

Proof By Lemma E.22, we have with probability at least 0.95 that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{Ax}\|_1 = (1 \pm O(\varepsilon))n\mu_{\mathbf{x}}$$

and with probability at least 0.95 that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_1 = \frac{n}{r} \cdot (1 \pm O(\varepsilon))r\mu_{\mathbf{x}} = (1 \pm O(\varepsilon))n\mu_{\mathbf{x}}.$$

Combining these two bounds, we have that with probability at least 0.9, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_1 = (1 \pm O(\varepsilon))\|\mathbf{Ax}\|_1.$$

■

E.6. Lower bound

In this section, we work towards proving a lower bound for a general class of random matrices with each column drawn i.i.d. from a different distribution. When specialized to our i.i.d. matrices from the above, we obtain nearly tight bounds. We do not have a single general theorem, but rather a number of different possible arguments that give better bounds depending on the underlying distribution. This can be shown to approximately recover the result of (Wang and Woodruff, 2019, Theorem 1.1), and additional yields new bounds, for example, the following tight results:

Theorem E.24 *Let $\log n \leqslant O(d)$ and let $\mathbf{S}$ be a $r \times n$ matrix such that*

$$\Pr_{\mathbf{A}\sim \mathrm{Cauchy}^{n\times d}}(\|\mathbf{A}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\|_1 \leqslant \kappa\|\mathbf{A}\|_1) \geqslant \frac{99}{100}$$

Then,

$$\kappa = \Omega\left(\frac{\log n}{\log r}\right)$$

Theorem E.25 *Let $\mathbf{S}$ be a $r \times n$ matrix such that*

$$\Pr_{\mathbf{A}\sim \mathcal{D}^{n\times d}}(\|\mathbf{A}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\|_1 \leqslant \kappa\|\mathbf{A}\|_1) \geqslant \frac{99}{100}$$

where $\mathcal{D}$ is a p-stable distribution. Then,

$$\kappa = \Omega\left(\frac{d^{1/p}}{r^{1-1/p}}\right)$$

Definition E.26 *Let $\mathcal{D}_j$ for $j \in [d]$ be distributions and consider the distribution $\mathcal{D}_{\mathbf{A}}$ over $n \times d$ matrices $\mathbf{A}$ that draws column j from $\mathcal{D}_j^n$. Let*

$$M_j \coloneqq \operatorname*{median}_{\mathbf{u}\sim\mathcal{D}_j^n}\|\mathbf{u}\|_1.$$

We then define the distribution $\mathcal{D}_{\max}$ that draws entries as

$$\max_{j=1}^{d} \frac{|v_j|}{M_j}, v_j \sim \mathcal{D}_j$$

and let its cdf be $F_{\mathcal{D}_{\max}}$.

Throughout this section, let $\mathbf{S}$ be a $r \times n$ matrix such that

$$\Pr_{\mathbf{A}\sim\mathcal{D}_{\mathbf{A}}}(\|\mathbf{A}\|_1 \leqslant \|\mathbf{S}\mathbf{A}\|_1 \leqslant \kappa\|\mathbf{A}\|_1) \geqslant 1-\delta.$$

E.6.1. PRELIMINARY BOUNDS ON $\mathbf{S}$

Lemma E.27 *For every $i \in [n]$, we have that*

$$\|\mathbf{S}\mathbf{e}_i\|_1 \leqslant 2\kappa\frac{1+F^{-1}_{\mathcal{D}_{\max}}(4\delta)}{F^{-1}_{\mathcal{D}_{\max}}(4\delta)}$$

Proof Note that for every row i of $\mathbf{A}$, with probability at least 4δ, one of the d columns of $\mathbf{e}_i^\top\mathbf{A}$, say column $j \in [d]$, has absolute value at least

$$\frac{|\mathbf{e}_i^\top\mathbf{A}\mathbf{e}_j|}{M_j} \geqslant F^{-1}_{\mathcal{D}_{\max}}(4\delta) \iff |\mathbf{e}_i^\top\mathbf{A}\mathbf{e}_j| \geqslant F^{-1}_{\mathcal{D}_{\max}}(4\delta)M_j.$$

Independently, with probability at least $1/2$, the ℓ_1 norm of the rest of the entries of the column is at most

$$\sum_{i' \in [n] \setminus \{i\}} |\mathbf{e}_{i'}^\top \mathbf{A} \mathbf{e}_j| \leqslant M_j.$$

Then with probability 2δ both of these happen simultaneously, so that

$$\|\mathbf{A}\mathbf{e}_j\|_1 \leqslant (1 + F_{\mathcal{D}_{\max}}^{-1}(4\delta))M_j$$

Let this event be $\mathcal{E}_i$.

Now suppose for contradiction that there is some column $i \in [n]$ such that $\|\mathbf{S}\mathbf{e}_i\|_1 > 2\kappa / F_{\mathcal{D}_{\max}}^{-1}(4\delta)$. We then condition on $\mathcal{E}_i$. Then with probability at least $1/2$,

$$\begin{aligned}\|\mathbf{S}\mathbf{A}\mathbf{e}_j\|_1 &= \|\mathbf{S}\mathbf{e}_i(\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j) + \sum_{i' \neq i} \mathbf{S}\mathbf{e}_{i'}(\mathbf{e}_{i'}^\top \mathbf{A}\mathbf{e}_j)\|_1 \geqslant \frac{1}{2}\|\mathbf{S}\mathbf{e}_i(\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j)\|_1 \\ &> \frac{1}{2}\left(2\kappa \frac{1 + F_{\mathcal{D}_{\max}}^{-1}(4\delta)}{F_{\mathcal{D}_{\max}}^{-1}(4\delta)}\right) F_{\mathcal{D}_{\max}}^{-1}(4\delta) M_j = \kappa(1 + F_{\mathcal{D}_{\max}}^{-1}(4\delta))M_j \geqslant \kappa \|\mathbf{A}\mathbf{e}_j\|_1\end{aligned}$$

so $\mathbf{S}$ fails to sketch $\mathbf{A}$ with probability at least δ, which is a contradiction. ■

Lemma E.28 *Let $F_{\mathcal{D}_{\max}}^{-1}(4\delta) < x < 1$. Then there are at most*

$$\frac{1}{1 - F_{\mathcal{D}_{\max}}(x)}$$

columns of $\mathbf{S}$ with ℓ_1 norm more than

$$2\kappa \frac{1+x}{x}.$$

Proof Let

$$p \coloneqq \mathop{\mathbf{Pr}}_{X \sim \mathcal{D}_{\max}}\{X \geqslant x\} = 1 - F_{\mathcal{D}_{\max}}(x)$$

and suppose for contradiction that there are more than $1/p$ columns of $\mathbf{S}$ with ℓ_1 norm more than $2\kappa/x$. Note that for each row i of these $1/p$ rows, there is a p probability that one of the d columns of $\mathbf{e}_i^\top \mathbf{A}$, say column $j \in [d]$, has absolute value at least

$$\frac{|\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j|}{M_j} \geqslant x \iff |\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j| \geqslant xM_j.$$

Then the probability that one of the $1/p$ rows, say row i, has an entry of absolute value at least xM_j is at least

$$1 - (1-p)^{1/p} \geqslant 1 - e^{-1}.$$

Independently, with probability at least $1/2$, the ℓ_1 norm of the rest of the entries of this column is at most

$$\sum_{i' \in [i] \setminus \{i\}} |\mathbf{e}_{i'}^\top \mathbf{A}\mathbf{e}_j| \leqslant M_j.$$

Thus with probability at least $(1-e^{-1})/2$, both of these events happen simultaneously, so there is row $i \in [n]$ and a column $j \in [d]$ such that

$$\begin{cases} |\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j| & \geqslant xM_j \\ \|\mathbf{A}\mathbf{e}_j\|_1 & \leqslant (1+x)M_j \\ \|\mathbf{S}\mathbf{e}_i\| & \geqslant 2\kappa\frac{1+x}{x} \end{cases}.$$

We condition on this event. Then with probability at least $1/2$,

$$\|\mathbf{S}\mathbf{A}\mathbf{e}_j\|_1 = \|\mathbf{S}\mathbf{e}_i(\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j) + \sum_{i' \neq i} \mathbf{S}\mathbf{e}_{i'}(\mathbf{e}_{i'}^\top \mathbf{A}\mathbf{e}_j)\|_1 \geqslant \frac{1}{2}\|\mathbf{S}\mathbf{e}_i(\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j)\|_1$$
$$> \frac{1}{2}2\kappa\frac{1+x}{x} \cdot xM_j = \kappa(1+x)M_j \geqslant \kappa\|\mathbf{A}\mathbf{e}_j\|_1$$

so $\mathbf{S}$ fails to sketch $\mathbf{A}$ with probability at least $(1-e^{-1})/4 > \delta$, which is a contradiction. ■

E.6.2. DISTORTION LOWER BOUND

Fix any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_1 = 1$ and let $\mathbf{v} = \mathbf{A}\mathbf{x}$. Following Wang and Woodruff (2019), our strategy is to bound $\|\mathbf{S}\mathbf{v}\|_1$ from above in terms of κ, and then derive a lower bound on κ by bounding $\|\mathbf{S}\mathbf{v}\|_1$ below by $\|\mathbf{v}\|_1$.

Note that the bound of Lemma E.28 is useless when

$$\frac{1}{1-F_{\mathcal{D}_{\max}}(x)} \geqslant n \iff F_{\mathcal{D}_{\max}}(x) \geqslant 1-\frac{1}{n} \iff x \geqslant F_{\mathcal{D}_{\max}}^{-1}\left(1-\frac{1}{n}\right).$$

We thus set $\mathbf{S}^H$ to be the matrix formed by taking the columns of $\mathbf{S}$ with ℓ_1 norm at least

$$2\kappa\frac{1+F_{\mathcal{D}_{\max}}^{-1}(1-1/n)}{F_{\mathcal{D}_{\max}}^{-1}(1-1/n)}$$

and $\mathbf{S}^L$ to be the columns of $\mathbf{S}$ with ℓ_1 norm at most this, and individually bound $\mathbf{S}^H\mathbf{v}$ and $\mathbf{S}^L\mathbf{v}$.

Now consider the distribution $\mathcal{D}_{\mathbf{x}}$ with cdf $F_{\mathbf{x}}$ that draws its entries as $|\langle \mathbf{x}, \mathbf{w}\rangle|$ with $\mathbf{w} \sim \prod_{j=1}^d \mathcal{D}_j$. By a union bound, the largest absolute value entry in $\mathbf{v} = \mathbf{A}\mathbf{x}$ is at most $M_{\mathbf{x}} \coloneqq F_{\mathbf{x}}^{-1}(1-1/2n)$ with probability at least $1/2$. Let this event be

$$\mathcal{E} \coloneqq \{\|\mathbf{v}\|_\infty \leqslant M_{\mathbf{x}}\}.$$

Throughout this section, we condition on $\mathcal{E}$. We also define

$$F_{\mathbf{x},\wedge}(x) \coloneqq \frac{F_{\mathbf{x}}(x)\mathbb{1}(x \leqslant M_{\mathbf{x}})}{\mathbf{Pr}(\mathcal{E})}$$

to be the conditional cdf of the capped version of $\mathbf{v}$.

E.6.3. BOUNDING THE HIGH-NORM COLUMNS OF $\mathbf{S}$

We first bound $\|\mathbf{S}^H\mathbf{v}\|_1$. We will need the following simple lemma:

Lemma E.29 *Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ be vectors with nonnegative entries and unit ℓ_1 norm. Let $\mathbf{P}$ be a uniformly random permutation matrix. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{P}}\langle \mathbf{u}, \mathbf{P}\mathbf{v}\rangle = \frac{1}{n}.$$

Proof We have that

$$\mathop{\mathbf{E}}_{\mathbf{P}}\langle \mathbf{u}, \mathbf{P}\mathbf{v}\rangle = \sum_{i=1}^n u_i \mathop{\mathbf{E}}_{\mathbf{P}}(\mathbf{e}_i^\top \mathbf{P}\mathbf{v}) = \sum_{i=1}^n u_i \sum_{j=1}^n \frac{v_j}{n} = \frac{1}{n}.$$

■

The main result for this section then is the following:

Lemma E.30 *Let*

$$L_{\max} \coloneqq F^{-1}_{\mathcal{D}_{\max}}\left(1 - \frac{1}{n}\right)$$
$$L_{\min} \coloneqq F^{-1}_{\mathcal{D}_{\max}}(4\delta)$$

Then,

$$\mathbf{Pr}\left\{\left\|\mathbf{S}^H \frac{\mathbf{v}}{\|\mathbf{v}\|_1}\right\|_1 \leqslant 400\frac{\kappa}{n}\sum_{k=\log_2 L_{\min}}^{\log_2 L_{\max}} \frac{1+2^{k-1}}{2^k(1 - F_{\mathcal{D}_{\max}}(2^k))}\right\} \geqslant \frac{99}{100}.$$

Proof Note that i.i.d. distributions are permutation invariant. We first fix the entries of $\mathbf{v}$, which fixes $\|\mathbf{v}\|_1$, but not the permutation of the entries. Now by Lemma E.29, we have

$$\mathop{\mathbf{E}}_{\mathbf{P}}\left\|\mathbf{S}^H\mathbf{P}\frac{\mathbf{v}}{\|\mathbf{v}\|_1}\right\|_1 = \sum_{i=1}^r \|\mathbf{e}_i^\top \mathbf{S}^H\|_1 \mathop{\mathbf{E}}_{\mathbf{P}}\left|\left\langle \frac{\mathbf{e}^\top\mathbf{S}^H}{\|\mathbf{e}_i^\top\mathbf{S}^H\|_1}, \mathbf{P}\frac{\mathbf{v}}{\|\mathbf{v}\|_1}\right\rangle\right| \leqslant \sum_{i=1}^r \|\mathbf{e}_i^\top\mathbf{S}^H\|_1 \frac{1}{n} = \frac{1}{n}\sum_{j=1}^n \|\mathbf{S}^H\mathbf{e}_j\|_1.$$

By Lemma E.28, we have that for each integer k between $\log_2 L_{\min}$ and $\log_2 L_{\max}$, there are at most

$$\frac{1}{1 - F_{\mathcal{D}_{\max}}(2^k)}$$

columns of $\mathbf{S}$ with ℓ_1 norm more than $2\kappa(1 + 1/2^k)$. Thus, there are at most $(1 - F_{\mathcal{D}_{\max}}(2^k))^{-1}$ columns with ℓ_1 norm in $[2\kappa(1+1/2^k), 2\kappa(1+1/2^{k-1})]$. Then, summing over the bounds over these intervals, we have that

$$\sum_{j=1}^n \|\mathbf{S}^H\mathbf{e}_j\|_1 \leqslant \sum_{k=\log_2 L_{\min}}^{\log_2 L_{\max}} 2\kappa\left(1 + \frac{1}{2^{k-1}}\right)\frac{1}{1 - F_{\mathcal{D}_{\max}}(2^k)} = \kappa \sum_{k=\log_2 L_{\min}}^{\log_2 L_{\max}} \frac{4(1+2^{k-1})}{2^k(1 - F_{\mathcal{D}_{\max}}(2^k))}.$$

Chaining together the inequalities gives the bound

$$\mathop{\mathbf{E}}_{\mathbf{P}}\left\|\mathbf{S}^H\mathbf{P}\frac{\mathbf{v}}{\|\mathbf{v}\|_1}\right\|_1 \leqslant \frac{\kappa}{n}\sum_{k=\log_2 L_{\min}}^{\log_2 L_{\max}} \frac{4(1+2^{k-1})}{2^k(1 - F_{\mathcal{D}_{\max}}(2^k))}.$$

We then conclude by Markov's inequality. ■

E.6.4. BOUNDING THE LOW-NORM COLUMNS OF $\mathbf{S}$

The idea for bounding $\|\mathbf{S}^L\mathbf{v}\|_1$ is that different arguments are needed for different level sets of $\mathbf{v}$, depending on how "spiky" it is. That is, a relatively flat level $\mathbf{v}_k$ should benefit from the sign cancellations in the product $\mathbf{e}_i^\top \mathbf{S}^L \mathbf{v}_k$, while a very spiky vector such as standard basis vectors should just apply the triangle inequality and bound only the few columns of $\mathbf{S}^L$ that it touches. This idea is formalized in the following lemma.

Lemma E.31 *Let $\mathbf{S} \in \mathbb{R}^{r\times n}$ be a fixed matrix such that $\|\mathbf{S}\mathbf{e}_i\|_1 \leqslant 1$ for each $i \in [n]$, and let $\mathbf{w} \in \mathbb{R}^n$ be a vector with entries drawn i.i.d. from a distribution with $\mathbf{E}\, w_i = 0$ and $\sigma^2 \coloneqq \mathbf{E}\, w_i^2 < \infty$. Let $\mu \coloneqq \mathbf{E}|w_i|$. Then,*

$$\mathbf{E}\|\mathbf{S}\mathbf{w}\|_1 \leqslant \min\{\mu n, C\sigma\sqrt{rn}\}$$

for an absolute constant C.

Proof For the first term of the min, we can simply use the triangle inequality to obtain

$$\mathbf{E}\|\mathbf{S}\mathbf{w}\|_1 \leqslant \sum_{i=1}^n \|\mathbf{S}\mathbf{e}_i\|_1\, \mathbf{E}|w_i| = \mu n.$$

For the second term, we first apply Jensen's inequality to get

$$\mathbf{E}\|\mathbf{S}\mathbf{w}\|_1 = \sum_{i=1}^r \mathbf{E}|\mathbf{e}_i^\top \mathbf{S}\mathbf{w}| \leqslant C\sum_{i=1}^r \left(\sum_{j=1}^n (\mathbf{e}_i^\top \mathbf{S}\mathbf{e}_j)^2\, \mathbf{E}\, w_j^2\right)^{1/2} = C\sigma \sum_{i=1}^r \|\mathbf{e}_i^\top \mathbf{S}\|_2$$

for some absolute constant C. We then finish by an application of Cauchy-Schwarz, switching from row-wise sums to column-wise sums, and bounding the ℓ_2 norm by the ℓ_1 norm:

$$\sum_{i=1}^r \|\mathbf{e}_i^\top \mathbf{S}\|_2 \leqslant \sqrt{r}\left(\sum_{i=1}^r \|\mathbf{e}_i^\top \mathbf{S}\|_2^2\right)^{1/2} = \sqrt{r}\left(\sum_{j=1}^n \|\mathbf{S}\mathbf{e}_j\|_2^2\right)^{1/2} \leqslant \sqrt{r}\left(\sum_{j=1}^n \|\mathbf{S}\mathbf{e}_j\|_1^2\right)^{1/2} \leqslant \sqrt{rn}.$$

■

Now for intuition, in Lemma E.31, we roughly think of the distribution of w_i as being v_i if v_i belongs to a level set, and 0 otherwise. Then if p is the probability of being in a given level set, the first term is roughly pn while the second term is roughly $\sqrt{rpn}$, so the first bound is tighter when $p \leqslant r/n$ and the second bound is tighter when $p \geqslant r/n$.

This yields the following:

Corollary E.32 *let*

$$T \coloneqq F_{\mathbf{x},\wedge}^{-1}\left(1 - \frac{r}{n}\right)$$

and write $\mathbf{v} = \mathbf{v}_{\leqslant T} + \mathbf{v}_{>T}$, where $\mathbf{v}_{\leqslant T}$ takes the value of $\mathbf{v}$ on coordinates $i \in [n]$ where $|v_i| \leqslant T$ and 0 otherwise, and $\mathbf{v}_{>T}$ similarly takes the coordinates $i \in [n]$ of $\mathbf{v}$ such that $|v_i| > T$ and 0 otherwise. Then, $\mathbf{v}_{\leqslant T}$ is drawn i.i.d. from a distribution with second moment

$$\sigma_{\leqslant T}^2 \coloneqq \int_0^T x^2 f_{\mathbf{x},\wedge}(x)\, dx$$

while $\mathbf{v}_{>T}$ *is drawn i.i.d. from a distribution with expected absolute value*

$$\mu_{>T} \coloneqq \int_T^{M_{\mathbf{x}}} x f_{\mathbf{x},\wedge}(x)\, dx.$$

Applying Lemma E.31, we obtain the bound

$$\mathbf{E}\|\mathbf{S}^L\mathbf{v}\|_1 \leqslant \mathbf{E}\|\mathbf{S}^L\mathbf{v}_{\leqslant T}\|_1 + \mathbf{E}\|\mathbf{S}^L\mathbf{v}_{>T}\|_1 \leqslant C\kappa\frac{1+F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}{F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}(\sigma_{\leqslant T}\sqrt{rn}+\mu_{>T}n).$$

E.6.5. LOWER BOUNDS FOR SKETCHING I.I.D. p-STABLE MATRICES

We apply Corollary E.32 to prove Theorem E.24:

Proof [Proof of Theorem E.25] Note that when $\mathbf{A}$ is drawn as fully i.i.d. Cauchy variables, then

$$F^{-1}_{\mathcal{D}_{\max}}(1-1/n) = O\left(\frac{nd}{n\log n}\right) = O(d/\log n).$$

We now apply Corollary E.32 with $\mathbf{v} = \mathbf{A}\mathbf{e}_1$, a Cauchy vector. Then, $T = \Theta(n/r)$, $\sigma^2_{\leqslant T} = \Theta(n/r)$, and $\mu_{>T} = \Theta(\log r)$ which yields

$$\mathbf{E}\|\mathbf{S}^L\mathbf{v}\|_1 \leqslant O\left(\kappa\frac{1+d/\log n}{d/\log n}\left[\sqrt{\frac{n}{r}}\sqrt{rn} + n\log r\right]\right) = O(\kappa n\log r).$$

Then with constant probability, we have

$$\Omega(n\log n) \leqslant \|\mathbf{v}\|_1 \leqslant \|\mathbf{S}^L\mathbf{v}\|_1 \leqslant O(\kappa n\log r)$$

and thus

$$\kappa = \Omega\left(\frac{\log n}{\log r}\right),$$

as desired. ■

When we have a column drawn i.i.d. from a p-stable distribution, we have an alternative bound:

Lemma E.33 *Let* $\mathbf{v}$ *be drawn i.i.d. from a* p*-stable distribution for* $p \in (1,2)$*. If* $\mathbf{v}$ *is in the column space of* $\mathbf{S}$*, then*

$$2\kappa\frac{1+F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}{F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}r^{1-1/p}n^{1/p} = \Omega(n).$$

Proof Then, we have that

$$\Omega(n) \leqslant \|\mathbf{v}\|_1 \leqslant \|\mathbf{S}^L\mathbf{v}\|_1 = \sum_{i=1}^r |\mathbf{e}_i^\top\mathbf{S}^L\mathbf{v}| = \sum_{i=1}^r \|\mathbf{e}_i^\top\mathbf{S}^L\|_p|\mathcal{S}_i|$$

By linearity of expectation, the above sum has expectation $\sum_{i=1}^r O(\|\mathbf{e}_i^\top\mathbf{S}^L\|_p)$, and thus is at most a constant times this with probability at least $99/100$ by a Markov bound. Then, we proceed by

bounding

$$\begin{aligned}\sum_{i=1}^{r}\|\mathbf{e}_i^\top \mathbf{S}^L\|_p &\leqslant r^{1-1/p}\left(\sum_{i=1}^{r}\|\mathbf{e}_i^\top \mathbf{S}^L\|_p^p\right)^{1/p}\\ &= r^{1-1/p}\left(\sum_{j=1}^{n}\|\mathbf{S}^L\mathbf{e}_j\|_p^p\right)^{1/p} \leqslant r^{1-1/p}\left(\sum_{j=1}^{n}\|\mathbf{S}^L\mathbf{e}_j\|_1^p\right)^{1/p}\\ &\leqslant r^{1-1/p}\left(n\left(2\kappa\frac{1+F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}{F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}\right)^p\right)^{1/p} = 2\kappa\frac{1+F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}{F^{-1}_{\mathcal{D}_{\max}}(1-1/n)}r^{1-1/p}n^{1/p}.\end{aligned}$$

■

This gives a proof of Theorem E.25.

Proof [Proof of Theorem E.25] When $\mathbf{A}$ is drawn as fully i.i.d. p-stable variables, then

$$F^{-1}_{\mathcal{D}_{\max}}(1-1/n) = \Theta\left(\frac{(nd)^{1/p}}{n}\right)$$

so by Lemma E.33, the distortion bound from these columns is

$$\Omega(n) \leqslant \kappa\frac{n}{(nd)^{1/p}}r^{1-1/p}n^{1/p} \iff \kappa \geqslant \Omega\left(\frac{d^{1/p}}{r^{1-1/p}}\right).$$

■

Appendix F. Missing proofs from Section A

Proof [Proof of Lemma A.4] For each $i \in S$, sample i with probability p and place the result in a uniformly random hash bucket in $[r]$ if it was sampled. Let $\mathcal{E}_i$ denote the event where i is sampled and is hashed to a bucket with no other members of T. Let $C_1, C_2, \ldots, C_{|S|}$ denote the sequence of these independent random choices and let $f(C_1, C_2, \ldots, C_s)$ denote the number of hash buckets in $[r]$ that contains members $i \in S$ satisfying $\mathcal{E}_i$ at the end of the sampling and hashing process. Note that f is 1-Lipschitz, and that

$$\mathbf{E}\, f(C_1, C_2, \ldots, C_{|S|}) = \sum_{i\in S}\mathbf{Pr}(\mathcal{E}_i) = |S|p\Big(1-\frac{p}{r}\Big)^{|T|} \geqslant p|S|\left(1-\frac{p|T|}{r}\right) \geqslant (1-\varepsilon)p|S|.$$

Now consider the Doob martingale

$$Z_k \coloneqq \mathbf{E}\big[f_q(C_1, C_2, \ldots, C_{|S|}) \mid C_1, C_2, \ldots, C_k\big].$$

Note that the increments $Z_k - Z_{k-1}$ conditioned on $C_1, C_2, \ldots, C_{k-1}$ is simply the indicator variable of whether on choice C_k we sampled an entry and placed it in a new bucket or not. Then

$Z_k - Z_{k-1} = 1$ with probability at most p and thus $\mathbf{E}_{k-1}(Z_k - Z_{k-1})^2 \leqslant p$. Then by Freedman's inequality Freedman (1975),

$$\mathbf{Pr}\big(|Z_{|S|} - Z_0| \geqslant \varepsilon Z_0\big) \leqslant 2\exp\left(-\frac{1}{2}\frac{(\varepsilon(1-\varepsilon)p|S|)^2}{p|S| + \varepsilon(1-\varepsilon)p|S|/3}\right) \leqslant 2\exp\left(-\frac{\varepsilon^2}{12}p|S|\right).$$

■

Proof [Proof of Theorem A.5] We make minor modifications of Lemmas 2.10 and 2.12 in Wang and Woodruff (2019). Let $\{X_i\}_{i=1}^n$ be independent standard Cauchys.

Upper bound. Let $\mathcal{E}_i := \big\{|X_i| \leqslant r\log n(\log\log r)^{-1}\big\}$. Then,

$$\mathbf{Pr}(\mathcal{E}_i) = 1 - \frac{2}{\pi}\arctan\left(\frac{r\log r}{\log\log r}\right) \geqslant 1 - \frac{2}{\pi}\frac{\log\log r}{r\log r} \gg \frac{1}{1+\varepsilon}.$$

Let $\mathcal{E} = \bigcap_{i=1}^r \mathcal{E}_i$. Then,

$$\mathbf{E}(|X_i| \mid \mathcal{E}) = \mathbf{E}(|X_i| \mid \mathcal{E}_i) = \frac{1}{\mathbf{Pr}(\mathcal{E}_i)}\frac{1}{\pi}\log\left(1 + \left(\frac{r\log r}{\log\log r}\right)^2\right)$$

and thus by linearity of expectation,

$$\mu := \mathbf{E}\left(\sum_{i=1}^r |X_i| \mid \mathcal{E}\right) = \frac{1}{\mathbf{Pr}(\mathcal{E}_i)}\frac{r}{\pi}\log\left(1 + \left(\frac{r\log r}{\log\log r}\right)^2\right) \leqslant (1+\varepsilon)\frac{2}{\pi}r\log r.$$

Now by a Chernoff bound applied to the $|X_i|(\log\log r/r\log r) \in [0,1]$ conditioned on $\mathcal{E}$,

$$\mathbf{Pr}\left(\sum_{i=1}^r |X_i| \geqslant (1+\varepsilon)\mu \mid \mathcal{E}\right) \leqslant \exp\left(-\frac{\varepsilon^2\mu}{3}\frac{\log\log r}{\log r}\right) = \exp\big(-\Theta(\varepsilon^2\log\log r)\big)$$

so

$$\begin{aligned}\mathbf{Pr}\left(\sum_{i=1}^r |X_i| \leqslant (1+\varepsilon)\mu\right) &\geqslant \mathbf{Pr}\left(\sum_{i=1}^n |X_i| \leqslant (1+\varepsilon)\mu \mid \mathcal{E}\right)\mathbf{Pr}(\mathcal{E}) \\ &\geqslant \big(1 - \exp\big(-\Theta(\varepsilon^2\log\log r)\big)\big)\left(1 - \frac{2}{\pi}\frac{\log\log r}{r\log r}\right)^r \geqslant 1 - \frac{(3/\varepsilon)^d}{\delta}.\end{aligned}$$

Lower bound. Let $T = \frac{(3/\varepsilon)^d}{\delta}$. Note that by Taylor expansion, there is a $T' \geqslant 0$ such that for $t \geqslant T'$,

$$\mathbf{Pr}(|X_i| > t) \geqslant \frac{2}{\pi}t^{-1} + O(t^{-3}).$$

Now for $i \geqslant 0$ and $j \in [r]$, define the indicator

$$N_j^i := \begin{cases} 1 & \text{if } |X_i| > (1+\varepsilon)^i T' \\ 0 & \text{otherwise}\end{cases}$$

and $N^i := \sum_{j\in[r]} N^i_j$. Note that by the Taylor expansion bound,

$$\mathbf{E}\, N^i \geqslant \frac{2r}{\pi}\frac{1}{(1+\varepsilon)^i T'}.$$

Then by Chernoff bounds,

$$\mathbf{Pr}\big(N^i \geqslant (1+\varepsilon)\,\mathbf{E}\, N^i\big) \leqslant \exp\left(-\frac{\varepsilon^2}{3}\frac{2r}{\pi}\frac{1}{(1+\varepsilon)^i T'}\right)$$

Now let $i_{\max}$ be the largest i such that

$$\exp\left(-\frac{\varepsilon^2}{3}\frac{2r}{\pi}\frac{1}{(1+\varepsilon)^i T'}\right) \leqslant \frac{1}{T}.$$

Then by a union bound over the first $i_{\max}$ level sets, $N^i \geqslant (2/\pi)r(1+\varepsilon)^i T'$ and thus with probability at least $1 - 1/T$,

$$\sum_{i=1}^{r}|X_i| \geqslant \sum_{i=0}^{i_{\max}}\frac{2}{\pi}r(1+\varepsilon)^i T' = \frac{2}{\pi}r\log\left(\frac{1}{T'}\frac{\varepsilon^2}{3}\frac{2r}{\pi}\frac{1}{\log T}\right) \geqslant (1-\varepsilon)\frac{2}{\pi}r\log r.$$

Net argument. Given the above concentration results, the rest of the argument proceeds as done in Wang and Woodruff (2019), using 1-stability of Cauchys and then a standard net argument. ■

Appendix G. No contraction bound

In this section, we prove a no contraction result for a generic M-sketch embedding with subsampling rates p_h as specified in Lemma G.2 and a hash bucket size of N_0 for the 0th level and N for the hth level for $h \in [h_{\max}]$ as specified in Definition G.4. This allows us to apply the results to both M-sketch with random and fixed boundaries, with varied branching factors and failure rates. Recall the definition of the M-sketch from Definition B.7.

Theorem G.1 (No contraction) *Let $\mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{y}\|_1 = 1$. Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Let $\mathbf{S}$ be drawn as an M-sketch matrix. Then with probability at least $1 - 6\delta$,*

$$\|\mathbf{SAy}\|_1 \geqslant (1 - 16\varepsilon)\|\mathbf{Ay}\|_1.$$

G.1. Essential weight classes

We first classify a small subset of weight classes of $\mathbf{y}$ that we need to preserve for at least a $(1-\varepsilon)$ approximation.

Lemma G.2 (Essential weight classes) *Let $\mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{y}\|_1 = 1$. Let $m_{\min}$ be a minimum class size parameter, let B be a branching factor parameter, and let ε be an accuracy parameter. Finally,*

let $p_h = p_0/B^{h-1}$ for $h \in [\log_B n]$ be sampling rates. Define

$$\begin{aligned} h_{\max} &\coloneqq \log_B n \\ q_{\max} &\coloneqq \log_2 \frac{n}{\varepsilon} \\ m_{\min} &\coloneqq \frac{12}{\varepsilon^2} \log \frac{4q_{\max}}{\delta} \\ M_{\geqslant} &\coloneqq \log_2 \frac{B}{\varepsilon} \\ M_{<} &\coloneqq \log_2 \frac{m_{\min}}{p_0 \varepsilon} \end{aligned}$$

and weight classes

$$\begin{aligned} \hat{Q}_h &\coloneqq \{q \in [q_{\max}] : m_{\min} \leqslant p_h |W_q(\mathbf{y})| < B m_{\min}\} && h \in [h_{\max}] \\ Q_h &\coloneqq \left\{ q \in \hat{Q}_h : q \leqslant M_{\geqslant} + \min_{q \in \hat{Q}_h} q, \|W_q(\mathbf{y})\|_1 \geqslant \frac{\varepsilon}{q_{\max}} \right\} && h \in [h_{\max}] \\ Q_{<} &\coloneqq \left\{ q : |W_q(\mathbf{y})| < m_{\min}/p_0, q \leqslant M_{<}, \|W_q\|_1 \geqslant \frac{\varepsilon}{M_{<}} \right\} \\ Q^* &\coloneqq Q_{<} \cup \bigcup_{h \in [h_{\max}]} Q_h \end{aligned}$$

Then,

$$\sum_{q \in Q^*} \|W_q(\mathbf{y})\|_1 \geqslant 1 - 6\varepsilon$$

Remark G.3 *The $\hat{Q}_h$ are the weight classes for which the hth level is the smallest level at which we sample at least $m_{\min}$ elements of W_q in expectation, so that the mass is extremely concentrated. The Q_h are the weight classes that restrict $\hat{Q}_h$ to only as many levels as we need to preserve the mass of $\hat{Q}_h$ up to a $1 - \varepsilon$ factor. The set $Q_<$ specifies the subset of levels that are too small for concentration, but are needed to preserve the mass of $\mathbf{y}$ up to a $1 - \varepsilon$ factor. The set Q^* specifies the union of these essential weight classes needed for a $1 - \varepsilon$ approximation.*

Proof Note that

$$\sum_{q > q_{\max}} \|W_q\|_1 \leqslant \frac{\varepsilon}{n} \sum_{q > q_{\max}} |W_q|_1 \leqslant \varepsilon$$

so we restrict our attention to $q \in [q_{\max}]$. Note that every $q \in [q_{\max}]$ belongs in either exactly one class $\hat{Q}_h$, or $|W|_q < m_{\min}/p_0$. The total weight of weight classes with $|W_q| < m_{\min}/p_0$ and $q > M_<$ is at most

$$\sum_{q > M_<} |W_q| 2^{1-q} = 2 \frac{m_{\min}}{p_0} 2^{-M_<} \sum_{q > 0} 2^{-q} \leqslant 2 \frac{m_{\min}}{p_0} \frac{p_0 \varepsilon}{m_{\min}} = 2\varepsilon.$$

Furthermore, let $h \in [h_{\max}]$ and let $q_h^* \coloneqq \min_{q \in \hat{Q}_h} q$. Then the ratio of the total weight of classes in W_q with $q > M_{\geqslant} + q_h^*$ to $\|W_{q_h^*}\|_1$ is at most

$$\frac{1}{\|W_{q_h^*}\|_1} \sum_{q > M_{\geqslant} + q_h^*} 2^{1-q} \frac{B m_{\min}}{p_h} \leqslant \frac{1}{2^{-q_h^*} m_{\min}/p_h} \sum_{q > M_{\geqslant} + q_h^*} 2^{1-q} \frac{B m_{\min}}{p_h}$$
$$= 2B \sum_{q > M_{\geqslant}} 2^{-q} \leqslant 2B 2^{-M_{\geqslant}} \leqslant 2B \frac{\varepsilon}{B} = 2\varepsilon.$$

We thus have that

$$\sum_{h \in [h_{\max}]} \sum_{q > M_{\geqslant} + q_h^*} \|W_q\|_1 \leqslant \sum_{h \in [h_{\max}]} 2\varepsilon \|W_{q_h^*}\|_1 \leqslant 2\varepsilon.$$

Furthermore, the total weight of classes in W_q with $\|W_q\|_1 < \varepsilon/q_{\max}$ is at most

$$\sum_{q : \|W_q\|_1 < \varepsilon/q_{\max}} \|W_q\|_1 \leqslant q_{\max} \frac{\varepsilon}{q_{\max}} = \varepsilon.$$

We conclude by combining the above bounds. ∎

G.2. Approximate perfect hashing

Definition G.4 (Useful constants)

$$N_0' \geqslant \frac{1}{\delta} \frac{M_<}{p_0 \varepsilon} m_{\min} \left(1 + \frac{7}{6} \frac{2 \log(2M_</\delta)}{\varepsilon^2}\right)$$
$$N_0 \coloneqq 2N_0' \log N_0' \qquad \textit{Number of hash buckets at the } 0\textit{th level}$$
$$N' \geqslant \frac{B}{\varepsilon} m_{\min} \left(M_{\geqslant} + \frac{7}{6} \frac{2 q_{\max} \log(2q_{\max}/\delta)}{\varepsilon^2}\right)$$
$$N \coloneqq 2N' \log N' \qquad \textit{Number of hash buckets}$$

We allow the flexibility to choose the number of buckets N_0 and N to be larger if needed. The N_0 and N are chosen so that

$$\frac{M_<}{p_0} m_{\min} \left(1 + \frac{7}{6} \frac{2 \log(2N_0 M_</\delta)}{\varepsilon^2}\right) \leqslant \delta \varepsilon N_0$$

and

$$B m_{\min} \left(M_{\geqslant} + \frac{7}{6} \frac{2 q_{\max} \log(2N q_{\max}/\delta)}{\varepsilon^2}\right) \leqslant \varepsilon N.$$

Lemma G.5 (Concentration of sampled mass) *Suppose $p_h |W_q| \geqslant m_{\min}$. Then with probability at least $1 - \delta/q_{\max}$,*

$$\sum_{\mathbf{y}_i \in W_q} b_{i,h} = (1 \pm \varepsilon) p_h |W_q|$$
$$\sum_{\mathbf{y}_i \in W_q} |\mathbf{y}_i| b_{i,h} = (1 \pm \varepsilon) p_h \|W_q\|_1$$

Proof Let

$$X \coloneqq \sum_{\mathbf{y}_i \in W_q} b_{i,h}.$$

By the Chernoff bound,

$$\mathbf{Pr}(|X - \mathbf{E}\, X| \geqslant \varepsilon\, \mathbf{E}\, X) \leqslant 2\exp\left(-\frac{\varepsilon^2\, \mathbf{E}\, X}{3}\right) \leqslant \frac{\delta}{2q_{\max}}.$$

Similarly, let

$$Y \coloneqq \sum_{\mathbf{y}_i \in W_q} |\mathbf{y}_i| b_{i,h}.$$

Note that

$$\begin{aligned} \mathbf{E}\, Y &= p_h \|W_q\|_1 \geqslant 2^{-q} p_h |W_q| \\ |\mathbf{y}_i| b_{i,h} &\leqslant 2^{1-q} \\ \mathbf{Var}(|\mathbf{y}_i| b_{i,h}) &\leqslant p_h 2^{2-2q} \end{aligned}$$

so by Bernstein's inequality,

$$\begin{aligned} \mathbf{Pr}(|Y - \mathbf{E}\, Y| \geqslant \varepsilon\, \mathbf{E}\, Y) &\leqslant 2\exp\left(-\frac{1}{2}\frac{(\varepsilon\, \mathbf{E}\, Y)^2}{p_h 2^{2-2q}|W_q| + (\varepsilon\, \mathbf{E}\, Y)2^{1-q}/3}\right) \\ &\leqslant 2\exp\left(-\frac{1}{2}\frac{(\varepsilon 2^{-q} p_h |W_q|)^2}{p_h 2^{2-2q}|W_q| + (\varepsilon 2^{1-q} p_h |W_q|)2^{1-q}/3}\right) \\ &= 2\exp\left(-\frac{1}{8}\frac{p_h |W_q| \varepsilon^2}{1 + \varepsilon/3}\right) \leqslant 2\exp\left(-\frac{\varepsilon^2}{12} p_h |W_q|\right) \leqslant \frac{\delta}{2q_{\max}}. \end{aligned}$$

We conclude by a union bound over the two events. ■

The following lemma uses a standard balls and bins martingale argument (e.g., Lee (2016)) to show that most items are hashed uniquely.

Lemma G.6 (Approximately perfect hashing) *Let $h \in [h_{\max}]$ and let $Q \subseteq \{q : p_h |W_q| \geqslant m_{\min}\}$. Let $\hat{W} \subset \mathbf{y}$ contain $W_Q \coloneqq \bigcup_{q \in Q} W_q$. Let $p_h |\hat{W}| \leqslant \varepsilon N$ for some $\varepsilon \in (0, 1/2)$. Then with probability at least $1 - (3/2)|Q|\delta/q_{\max}$, every W_q has a $W_q^* \subset W_q$ that gets sampled and placed in a hash bucket with no other members of $\hat{W}$, and $|W_q^*| \geqslant (1 - 3\varepsilon) p_h |W_q|$ and $\|W_q^*\|_1 \geqslant (1 - 9\varepsilon) p_h \|W_q\|_1$.*

Proof We apply Lemma A.4 to see that with probability at least

$$1 - 2\exp\left(-\frac{\varepsilon^2}{12} p_h |W_q|\right) \leqslant 1 - \frac{\delta}{2q_{\max}},$$

there is a set $W_q^* \subseteq W_q$ of elements that are hashed to a bucket with no other element of $\hat{W}$ in it and of size $|W_q^*| \geqslant (1 - \varepsilon)^2 p_h |W_q| \geqslant (1 - 3\varepsilon) p_h |W_q|$ with probability at least $1 - \delta/2q_{\max}$. We condition on this event.

By Lemma G.5, with probability at least $1 - \delta/q_{\max}$, we sample $(1 \pm \varepsilon)p_h|W_q|$ elements with mass $(1 \pm \varepsilon)p_h\|W_q\|_1$. Note then that there are at most $4\varepsilon p_h|W_q|$ sampled elements that do not belong W_q^*. The mass of these elements is at most

$$4\varepsilon p_h|W_q|2^{1-q} \leqslant 8\varepsilon p_h\|W_q\|_1.$$

Thus,

$$\|W_q^*\|_1 \geqslant (1-\varepsilon)p_h\|W_q\|_1 - 8\varepsilon p_h\|W_q\|_1 = (1-9\varepsilon)p_h\|W_q\|_1.$$

We conclude by a union bound over the weight classes Q. ■

G.3. Preserving weight classes

Definition G.7

$$\tau_0 \coloneqq \frac{p_0\varepsilon}{2M_< m_{\min}} \qquad \textit{Size of a relatively large element at 0th level}$$

$$T_0 \coloneqq \frac{6}{7}\frac{\varepsilon\tau_0}{\log(2N_0M_</\delta)} \qquad \textit{Size of a relatively small element at 0th level}$$

$$\tau_h \coloneqq \frac{p_h\varepsilon}{2q_{\max}Bm_{\min}} \qquad \textit{Size of a relatively large element}$$

$$T_h \coloneqq \frac{6}{7}\frac{\varepsilon\tau_h}{\log(2Nq_{\max}/\delta)} \qquad \textit{Size of a relatively small element}$$

Definition G.8 (Large elements)

$$Q_{<,0} \coloneqq \left\{q : q \leqslant \log_2\frac{1}{T_0}\right\}$$

$$Q_{<,h} \coloneqq \left\{q : q \leqslant \log_2\frac{1}{T_h}\right\}$$

The weight class $Q_{<,h}$ is the set of relatively large elements at the hth level of sampling.

We directly recall the following Lemma 3.3 from Clarkson and Woodruff (2015).

Lemma G.9 *Let $h \in [h_{\max}]$, $\bar{W} \subset \mathbf{y}$, $T \geqslant \|\bar{W}\|_\infty$, and $\delta' \in (0,1)$. If*

$$N \geqslant \frac{6\|\bar{W}\|_1}{T\log(N/\delta)},$$

then

$$\mathbf{Pr}\left(\max_{k\in[N]}\|L_{h,k}\cap\bar{W}\|_1 \leqslant \frac{7}{6}T\log(N/\delta)\right) \geqslant 1-\delta'.$$

G.3.1. PRESERVING WEIGHT CLASSES IN Q_h

Lemma G.10 *Let $h \in [h_{\max}]$, $q \in Q_h$. Then*

$$|\mathbf{y}_i| \geqslant \tau_h = \frac{p_h\varepsilon}{2q_{\max}Bm_{\min}}.$$

Proof By the definition of Q_h, we have that $m_{\min} \leqslant p_h|W_q| \leqslant Bm_{\min}$ and $\|W_q\|_1 \geqslant \frac{\varepsilon}{q_{\max}}$. We then have that

$$|W_q|2^{1-q} \geqslant \|W_q\|_1 \geqslant \frac{\varepsilon}{q_{\max}}.$$

Then for any $\mathbf{y}_i \in W_q$,

$$|\mathbf{y}_i| \geqslant 2^{-q} \geqslant \frac{\varepsilon}{2|W_q|q_{\max}} \geqslant \frac{p_h\varepsilon}{2q_{\max}Bm_{\min}}.$$

■

Lemma G.11 *Let $h \in [h_{\max}]$ and let $L_h(\mathbf{y}_i)$ denote the multiset of elements in the hash bucket in the hth level containing $\mathbf{y}_i$. Then with probability at least $1 - 2|Q_h|\delta/q_{\max}$, for all $q \in Q_h$, we sample a set $W_q^* \subseteq W_q$ such that*

$$\|W_q^*\|_1 \geqslant (1 - 9\varepsilon)p_h\|W_q\|_1$$

and for every $\mathbf{y}_i \in W_q^*$,

$$\left|\sum_{\mathbf{y}_j \in L_h(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j\right| \geqslant (1-\varepsilon)|\mathbf{y}_i|.$$

Proof Let $\hat{W} = W_{Q_h} \cup W_{Q_{<,h}}$. Then by our choice of N,

$$|\hat{W}| \leqslant |W_{Q_h}| + |W_{Q_{<,h}}| \leqslant M_{\geqslant}\frac{Bm_{\min}}{p_h} + \frac{1}{T_h} = \frac{B}{p_h}m_{\min}\left(M_{\geqslant} + \frac{7}{6}\frac{2q_{\max}\log(2Nq_{\max}/\delta)}{\varepsilon^2}\right) \leqslant \frac{\varepsilon N}{p_h}.$$

Then by Lemma G.6, with probability at least $1 - (3/2)|Q_h|/q_{\max}$, for each $q \in Q_h$, there is a set of sampled elements $W_q^* \subseteq W_q$ that get hashed to a bucket with no other members of $\hat{W}$, and $\|W_q^*\|_1 \geqslant (1 - 9\varepsilon)p_h\|W_q\|_1$.

Note that for each $q \in Q_h$ and $\mathbf{y}_i \in W_q^*$, the absolute value of the largest element in $L_h(\mathbf{y}_i)$ not equal to $\mathbf{y}_i$ is at most T_h, since we have hashed the elements of $W_{Q_{<,h}}$ to other buckets. Then by Lemma G.9, the ℓ_1 mass of elements that are at most T_h in all hash buckets are at most

$$\|L_h(\mathbf{y}_i) \setminus \{\mathbf{y}_i\}\|_1 \leqslant \frac{7}{6}T_h\log(2Nq_{\max}/\delta) = \varepsilon\tau_h$$

with probability at least $1 - \delta/2q_{\max}$. By a union bound over $q \in Q_h$, this is true for all $\mathbf{y}_i \in W_q^*$ for $q \in Q_h$ with probability at least $1 - 2|Q_h|\delta/q_{\max}$.

Recall from Lemma G.10 that $|\mathbf{y}_i| \geqslant \tau_h$ for all $\mathbf{y}_i \in W_q$ with $q \in Q_h$. Note then that the mass of this hash bucket is at least

$$\left|\sum_{\mathbf{y}_j \in L_h(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j\right| \geqslant |\mathbf{y}_i| - \|L_h(\mathbf{y}_i) \setminus \{\mathbf{y}_i\}\|_1 \geqslant |\mathbf{y}_i| - \varepsilon\tau_h \geqslant (1-\varepsilon)|\mathbf{y}_i|$$

which is the desired bound. Thus overall, the total success probability is at least $1 - 2|Q_h|\delta/q_{\max}$. ■

Lemma G.12 *Let $h \in [h_{\max}]$. Then with probability at least $1 - 2|Q_h|\delta/q_{\max}$, we have that*

$$\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}\|_1 \geqslant (1 - 10\varepsilon) \sum_{q \in Q_h} \|W_q\|_1.$$

Proof Taking a sum over $q \in Q_h$ and $\mathbf{y}_i \in W_q^*$, we find that

$$\begin{aligned}
\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}\|_1 &\geqslant \frac{1}{p_h} \sum_{q \in Q_h} \sum_{\mathbf{y}_i \in W_q} b_{i,h} \left| \sum_{\mathbf{y}_j \in L_h(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j \right| && \text{Looking only at rows in } Q_h \\
&\geqslant \frac{1}{p_h} \sum_{q \in Q_h} \sum_{\mathbf{y}_i \in W_q^*} \left| \sum_{\mathbf{y}_j \in L_h(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j \right| && \text{Looking only at good sampled elements } W_q^* \\
&\geqslant \frac{1}{p_h} \sum_{q \in Q_h} \sum_{\mathbf{y}_i \in W_q^*} (1 - \varepsilon)|\mathbf{y}_i| && \text{Lemma G.11} \\
&\geqslant (1 - \varepsilon)\frac{1}{p_h} \sum_{q \in Q_h} \|W_q^*\|_1 \\
&\geqslant (1 - \varepsilon)(1 - 9\varepsilon)\frac{1}{p_h} \sum_{q \in Q_h} p_h \|W_q\|_1 && \text{Lemma G.11} \\
&\geqslant (1 - 10\varepsilon) \sum_{q \in Q_h} \|W_q\|_1
\end{aligned}$$

which is the desired bound. The failure probability is the same as from Lemma G.11. ■

G.3.2. PRESERVING WEIGHT CLASSES IN $Q_<$

With essentially the exact same proofs as in the above section, we have the following analogues of Lemmas G.10, G.11, and G.12.

Lemma G.13 *Let $q \in Q_<$. Then*

$$|\mathbf{y}_i| \geqslant \tau_0 = \frac{\varepsilon p_0}{M_< m_{\min}}.$$

Lemma G.14 *Let $Q = \{q \in Q_< : |W_q| \geqslant m_{\min}\}$. Let $L_0(\mathbf{y}_i)$ denote the multiset of elements in the hash bucket in the 0th level containing $\mathbf{y}_i$. Then with probability at least $1 - 2|Q|\delta/M_<$, for all $q \in Q$, there is a set $W_q^* \subseteq W_q$ such that W_q^* is hashed to a different bucket than $W_{Q_<,0} \supset W_{Q_<}$,*

$$\|W_q^*\|_1 \geqslant (1 - 9\varepsilon)p_h\|W_q\|_1,$$

and for every $\mathbf{y}_i \in W_q^*$,

$$\left| \sum_{\mathbf{y}_j \in L_0(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j \right| \geqslant (1 - \varepsilon)|\mathbf{y}_i|.$$

Lemma G.15 *Let $Q = \{q \in Q_< : |W_q| \geqslant m_{\min}\}$. Then with probability at least $1 - 2|Q|\delta/M_<$,*

$$\|\mathbf{C}^{(0)}\mathbf{y}\|_1 \geqslant (1 - 10\varepsilon) \sum_{q \in Q} \|W_q\|_1.$$

It thus remains to handle the case of $\{q \in Q_< : |W_q| < m_{\min}\}$. For these small level sets, we can perfectly hash these into separate buckets from all the entries in $Q_{<,0}$.

Lemma G.16 *Let $Q = \{q \in Q_< : |W_q| < m_{\min}\}$. Let $L_0(\mathbf{y}_i)$ denote the multiset of elements in the hash bucket in the 0th level containing $\mathbf{y}_i$. With probability at least $1 - 2\delta$, every member of W_Q is hashed to a different bucket than $W_{Q_{<,0}} \supset W_{Q_<}$, and we have for every $\mathbf{y}_i \in W_Q$ that*

$$\left| \sum_{\mathbf{y}_j \in L_0(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j \right| \geqslant (1 - \varepsilon)|\mathbf{y}_i|.$$

Proof Note that

$$N_0 \geqslant \frac{1}{\delta}|W_Q||W_{Q_{<,0}}|.$$

Then for every $(\mathbf{y}_i, \mathbf{y}_j) \in W_Q \times W_{Q_{<,0}}$, there is a $\delta/|W_Q||W_{Q_{<,0}}|$ probability that $\mathbf{y}_i$ and $\mathbf{y}_j$ get hashed to the same location. By a union bound, none of these pairs are hashed to to the same location with probability at least $1 - \delta$. Then by Lemma G.9, the ℓ_1 mass of elements that are at most T_0 in all hash buckets are at most

$$\|L_0(\mathbf{y}_i) \setminus \{\mathbf{y}_i\}\|_1 \leqslant \frac{7}{6} T_0 \log(2N_0 M_</\delta) = \varepsilon\tau_0$$

with probability at least $1 - \delta/2M_<$. By a union bound over $q \in Q$, this is true for all $\mathbf{y}_i \in W_Q$ with probability at least $1 - |Q|\delta/2M_< \geqslant 1 - \delta$. Then,

$$\left| \sum_{\mathbf{y}_j \in L_0(\mathbf{y}_i)} \Lambda_j \mathbf{y}_j \right| \geqslant |\mathbf{y}_i| - \|L_0(\mathbf{y}_i) \setminus \{\mathbf{y}_i\}\|_1 \geqslant |\mathbf{y}_i| - \varepsilon\tau_0 \geqslant (1 - \varepsilon)|\mathbf{y}_i|$$

which is the desired bound. Thus overall, the failure probability is $1 - 2\delta$. ■

G.4. Proof of Theorem G.1

We finally gather the pieces from above.

Proof Proof of Theorem G.1 We union bound over the events and sum over the results of Lemmas G.15, G.16, and G.12, so that with probability at least $1 - 6\delta$,

$$\|\mathbf{SAy}\|_1 \geqslant (1 - 10\varepsilon) \sum_{q \in Q_< \cup \bigcup_{h \in [h_{\max}]} Q_h} \|W_q\|_1.$$

We conclude by chaining this inequality together with the result of Lemma G.2. ■

Appendix H. Missing proofs from Section C

Proof [Proof of Lemma C.4] By Hoeffding bounds, we have for each $j \in [d]$ that

$$\mathbf{Pr}\left(\left|\sum_{i=1}^{s} \varepsilon_i \mathbf{e}_j^\top \mathbf{x}_i\right| > \sqrt{\frac{s}{2}\log\frac{2d}{\delta}}\right) \leqslant 2\exp\left(-\frac{2(\sqrt{(s/2)\log(2d/\delta)})^2}{s}\right) \leqslant \frac{\delta}{d}$$

Then by a union bound over the d choices of j, with probability at least $1-\delta$, the complement event of the above holds for every $j \in [d]$. Conditioned on this event, we have that

$$\left\|\sum_{i=1}^{s} \varepsilon_i \mathbf{x}_i\right\|_1 \leqslant \sum_{j=1}^{d}\left|\sum_{i=1}^{s} \varepsilon_i \mathbf{e}_j^\top \mathbf{x}_i\right| \leqslant d\sqrt{\frac{1}{2}\log\frac{2d}{\delta}}\sqrt{s}$$

as desired. ■

Proof [Proof of Lemma C.7] For each $i \in [r]$ and $j \in [d]$, by the 1-stability of Cauchy variables,

$$\mathbf{e}_i^\top \mathbf{S}\mathbf{A}\mathbf{e}_j \overset{d}{=} \|\mathbf{e}_i^\top \mathbf{S}\|_1 \mathcal{C}_{i,j}$$

where $\mathcal{C}_{i,j}$ are drawn as standard Cauchy variables, and are independent for distinct j. Now note that $|\mathcal{C}_{i,j}| \leqslant O(rd)$ with probability at least $1-(100rd)^{-1}$ and thus by a union bound, $\max_{i\in[r],j\in[d]}|\mathcal{C}_{i,j}| \leqslant O(rd)$ with probability at least $1-1/400$. We condition on this event. Note then that the conditional expectation is at most

$$\mathbf{E}|\mathcal{C}_{i,j}| \leqslant O(\log(rd))$$

as shown in Indyk (2006a). Then,

$$\mathbf{E}\|\mathbf{S}\mathbf{A}\|_1 = \sum_{j=1}^{d}\sum_{i=1}^{r}\mathbf{E}|\mathbf{e}_i^\top \mathbf{S}\mathbf{A}\mathbf{e}_j| = \sum_{j=1}^{d}\sum_{i=1}^{r}\|\mathbf{e}_i^\top \mathbf{S}\|_1\,\mathbf{E}|\mathcal{C}_{i,j}| = O(d\log(rd))\|\mathbf{S}\|_1$$

so a Markov bound and a union bound with the earlier event shows that

$$\mathbf{Pr}(\|\mathbf{S}\mathbf{A}\|_1 \leqslant O(d\log(rd))\|\mathbf{S}\|_1) \geqslant 1 - \frac{1}{200}.$$

For the lower bound, let $\hat{\mathcal{C}}_{i,j}$ be the truncation of $\hat{\mathcal{C}}_{i,j}$ at d, i.e.,

$$\hat{\mathcal{C}}_{i,j} = \begin{cases} \mathcal{C}_{i,j} & \text{if } |\mathcal{C}_{i,j}| \leqslant d \\ 0 & \text{otherwise} \end{cases}.$$

Note then that by (Indyk, 2006a, Lemma 6), $\mathbf{Var}(|\hat{C}_{i,j}|) = \Theta(d)$ so

$$\sigma^2 := \mathbf{Var}\left(\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\hat{\mathcal{C}}_{i,j}|\right) = \sum_{j=1}^{d}\Theta(d)\|\mathbf{e}_i^\top \mathbf{S}\|_1^2 = \Theta(d^2)\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1^2$$

and

$$\mu := \mathbf{E}\left(\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\hat{\mathcal{C}}_{i,j}|\right) = \Theta(d\log d)\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1.$$

Then by Chebyshev's inequality,

$$\mathbf{Pr}\left(\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\hat{\mathcal{C}}_{i,j}| - \mu \leqslant \Theta(\log d)\sigma\right) \leqslant \frac{1}{400}.$$

Thus, with probability at least $1 - 1/400$,

$$\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\mathcal{C}_{i,j}| \geqslant \sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\hat{\mathcal{C}}_{i,j}| \geqslant \gamma\|\mathbf{e}_i^\top \mathbf{S}\|_1$$

for $\gamma = \Omega(d\log d)$. Let $\mathcal{E}_i$ denote the above event, so that $\mathbf{Pr}(\mathcal{E}_i) \geqslant 1 - 1/400$. Then,

$$\mathbf{E}\left(\sum_{i=1}^{r}\mathbb{1}(\neg\mathcal{E}_i)\left[\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\mathcal{C}_{i,j}|\right]\right) \leqslant \sum_{i=1}^{r}\frac{1}{400}\gamma\|\mathbf{e}_i^\top \mathbf{S}\|_1$$

so by Markov's inequality, with probability at least $1 - 1/200$,

$$\sum_{i=1}^{r}\mathbb{1}(\neg\mathcal{E}_i)\left[\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\mathcal{C}_{i,j}|\right] \leqslant \frac{\gamma}{2}\sum_{i=1}^{r}\|\mathbf{e}_i^\top \mathbf{S}\|_1 = \frac{\gamma}{2}\|\mathbf{S}\|_1.$$

Then, conditioning on this event,

$$\|\mathbf{SA}\|_1 \geqslant \sum_{i=1}^{r}\left[\sum_{j=1}^{d}\|\mathbf{e}_i^\top \mathbf{S}\|_1|\mathcal{C}_{i,j}|\right](1 - \mathbb{1}(\neg\mathcal{E}_i)) \geqslant \gamma\sum_{i=1}^{r}\|\mathbf{e}_i^\top \mathbf{S}\|_1 - \frac{\gamma}{2}\|\mathbf{S}\|_1 = \frac{\gamma}{2}\|\mathbf{S}\|_1$$

as desired. ■

Appendix I. Missing proofs from Section E

Proof [Proof of Lemma E.6] Because $\mathbf{Pr}(|X| \leqslant T) = \Theta(1)$ for T large enough,

$$\mathop{\mathbf{E}}_{X\sim\mathsf{trunc}_T(\mathcal{D})}|X| = \Theta(1)\mathop{\mathbf{E}}_{X\sim\mathcal{D}}(|X| \mid |X| \leqslant T), \qquad \mathop{\mathbf{E}}_{X\sim\mathsf{trunc}_T(\mathcal{D})}X^2 = \Theta(1)\mathop{\mathbf{E}}_{X\sim\mathcal{D}}(X^2 \mid |X| \leqslant T)$$

By the layer cake theorem,

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim\mathcal{D}}(|X| \mid |X| \leqslant T) &= \int_0^\infty \mathbf{Pr}(|X| > x \mid |X| \leqslant T)\,dx \\
&= \int_0^T \frac{\mathbf{Pr}(x < |X| \leqslant T)}{\mathbf{Pr}(|X| \leqslant T)} \\
&= \frac{1}{\mathbf{Pr}(|X| \leqslant T)}\int_0^T \mathbf{Pr}(|X| > x) - \mathbf{Pr}(|X| > T)\,dx \\
&= \Theta(1)\int_0^T \Theta(x^{-p})\,dx
\end{aligned}$$

and similarly,

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}}(X^2 \mid |X| \leqslant T) = \Theta(1)\int_0^T x\,\mathbf{Pr}(|X| > x)\,dx = \Theta(1)\int_0^T \Theta(x^{1-p})\,dx.$$

Solving the simple integrals yields the desired results. ■

Proof [Proof of Theorem E.8] The distortion upper bound is just Lemma E.2.

Mass of small entries. Let $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ as in Definition E.7, with $T = O\big((nd^2\log d/r)^{1/p}\big)$. Then, $\mathbf{A}^L \sim \mathsf{trunc}_T(\mathcal{D})^{n\times d}$ where by Lemma E.6, the first two moments of each entry are

$$\mu = \Theta(T^{1-p}), \qquad \sigma = \Theta(T^{2-p}).$$

Then by Bernstein's inequality,

$$\begin{aligned}-\log\mathbf{Pr}\big(\|\mathbf{A}^L\mathbf{e}_j\|_1 \geqslant 2\mu n\big) &\geqslant \frac{1}{2}\frac{(\mu n)^2}{\sigma^2 n + \mu n T/3}\\ &= \frac{\Omega((T^{1-p}n)^2)}{O(T^{2-p})n + O(T^{1-p})nT} = \Omega(nT^{-p}) = \Omega\left(\frac{r}{d^2\log d}\right) = \Omega(\log d)\end{aligned}$$

Thus, $\mathbf{Pr}(\|\mathbf{A}^L\mathbf{e}_j\|_1 \leqslant 2\mu n) \geqslant 1 - 1/\operatorname{poly}(d)$ so by a union bound over the d columns, this event simultaneously holds for all d columns with probability at least $1 - 1/\operatorname{poly}(d)$. Conditioned on this event, by the triangle inequality,

$$\|\mathbf{A}^L\mathbf{x}\|_1 \leqslant O\left(\frac{n^{1/p}}{(\log d)^{1/p-1}}\right)\|\mathbf{x}\|_1$$

for all $\mathbf{x}\in\mathbb{R}^d$.

Mass of large entries. Furthermore, let $\mathbf{B}'$ be the subset of rows of $\mathbf{A}^H$ given by Lemma E.11 that are hashed to locations without any other rows of $\mathbf{A}^H$. Recall also τ_1 and τ_2 from the lemma.

We first have that $\|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 = \Omega(\|\mathbf{A}^H\mathbf{x}\|_1)$ since the rows containing entries larger than τ_1 are perfectly hashed, while rows containing entries between τ_2 and τ_1 are preserved up to constant factors.

Let $\mathbf{B}' = \mathbf{B}'_{>T} + \mathbf{B}'_{\leqslant T}$ where $\mathbf{B}'_{>T}$ contains the entries of $\mathbf{B}'$ that have absolute value greater than T and $\mathbf{B}'_{\leqslant T}$ contains the rest of the entries. Note then that $\mathbf{B}'_{>T}$ has at most one nonzero entry per row, and $\mathbf{B}'_{\leqslant T}$ has at most $O(d\cdot r/d\log d) = O(r/\log d)$ nonzero entries and thus by Lemma

E.4, $\|\mathbf{B}'_{\leqslant T}\|_\infty \leqslant O(r^{1/p})$ with probability at least 0.99. We condition on this event. Then for all $\mathbf{x}$,

$$\begin{aligned}
\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 &\geqslant \|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 \\
&\geqslant \|\mathbf{S}\mathbf{B}'_{>T}\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 \\
&= \sum_{j=1}^{d} |\mathbf{x}_j| \|\mathbf{B}'_{>T}\mathbf{e}_j\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \mathbf{B}'_{>T}\mathbf{e}_j \text{ have disjoint support} \\
&\geqslant \sum_{j=1}^{d} |\mathbf{x}_j| \sum_{k=\log_2 \tau_1}^{\log_2 \tau_2} 2^k \Theta(n/2^{kp}) - \|\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \text{Lemmas E.11 and E.2} \\
&= \Omega((n/\log d)^{1/p} \log d)\|\mathbf{x}\|_1 - O(r)\|\mathbf{B}'_{\leqslant T}\|_\infty \|\mathbf{x}\|_1 && \text{Hölder's inequality} \\
&= \Omega(n^{1/p}/(\log d)^{1/p-1})\|\mathbf{x}\|_1 - O(r^{1+1/p})\|\mathbf{x}\|_1 \\
&= \Omega(n^{1/p}/(\log d)^{1/p-1})\|\mathbf{x}\|_1.
\end{aligned}$$

Conclusion. On the other hand, by Lemma E.14, the mass of the $O(r/d \log d)$ rows that are hashed together with the rows of $\mathbf{A}^H$ have mass at most

$$O\left(\frac{1}{\sqrt{\log d}} \frac{n^{1/p}}{(r/d^2 \log d)^{1/p-1}}\right)\|\mathbf{x}\|_1 = o\Big(n^{1/p}/(\log d)^{1/p-1}\Big)\|\mathbf{x}\|_1.$$

Then,

$$\frac{1}{\kappa} \geqslant \frac{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{A}\mathbf{x}\|_1} \geqslant \frac{\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{C}_1\mathbf{x}\|_1}{\|\mathbf{A}^H\mathbf{x}\|_1 + \|\mathbf{A}^L\mathbf{x}\|_1} \geqslant \frac{\Omega(\|\mathbf{A}^H\mathbf{x}\|_1 + n^{1/p}/(\log d)^{1/p-1})\|\mathbf{x}\|_1}{O(\|\mathbf{A}^H\mathbf{x}\|_1 + n^{1/p}/(\log d)^{1/p-1}))\|\mathbf{x}\|_1} \geqslant \Omega(1).$$

■

Proof [Proof of Lemma E.13] For a hash bucket $i \in [r]$ and $k \in [d]$, let

$$Y_{i,k} \coloneqq \left|\sum_{j:h(j)=i} \mathbf{e}_j^\top \mathbf{C}\mathbf{e}_k\right|$$

where h is the hash function for the CountSketch matrix $\mathbf{S}$. By Chernoff bounds and a union bound, there are $\Theta(n/r)$ rows $j \in [n]$ such that $h(j) = i$ for all buckets $i \in [r]$, with probability at least $1 - r\exp(-\Theta(n/r)) = 1 - o(1)$. Conditioned on this event, which is independent of the randomness of $\mathbf{C}$,

$$\begin{aligned}
\mathbf{E}\, Y_{i,k}^2 &= \sum_{j_1,j_2 \in h^{-1}(i)\times h^{-1}(i)} \mathbf{E}\Big[(\mathbf{e}_{j_1}^\top \mathbf{C}\mathbf{e}_k)(\mathbf{e}_{j_2}^\top \mathbf{C}\mathbf{e}_k)\Big] \\
&= \sum_{j:h(j)=i} \mathbf{E}\Big(\mathbf{e}_j^\top \mathbf{C}\mathbf{e}_k\Big)^2 = O\Big(\frac{n}{r}T^{2-p}\Big) = O\left((d^2 \log d)^{(2-p)/p}\Big(\frac{n}{r}\Big)^{2/p}\right)
\end{aligned}$$

by the second moment bound in Lemma E.6.

Now let S be the subset of rows of $\mathbf{S}'$. Then for each $k \in [d]$,

$$\mathbf{E}\left[\sum_{i\in S} Y_{i,k}\right] = \sum_{i\in S} \mathbf{E}\, Y_{i,k} \leqslant \sum_{i\in S} \sqrt{\mathbf{E}\, Y_{i,k}^2} = O\Big(r'(d^2\log d)^{1/p-1/2}(n/r)^{1/p}\Big)$$
$$\mathbf{Var}\left(\sum_{i\in S} Y_{i,k}\right) = \sum_{i\in S} \mathbf{Var}(Y_{i,k}) = O\Big(r'(d^2\log d)^{(2-p)/p}(n/r)^{2/p}\Big).$$

By Chebyshev's inequality,

$$\mathbf{Pr}\left(\sum_{i\in S} Y_{i,k} \leqslant \mathbf{E}\left[\sum_{i\in S} Y_{i,k}\right] + \lambda\sqrt{\mathbf{Var}\left(\sum_{i\in S} Y_{i,k}\right)}\right) \geqslant 1 - \frac{1}{\lambda}$$

which gives the desired result. ■

Proof [Proof of Lemma E.16] We compare $\mathcal{D}$ to a p-stable distribution $\mathcal{D}_p$. By (Nolan, 2018, Theorem 1.12), a p-stable distribution is a power law with index p. Then, there exist constants T and c such that for all $t \geqslant T$,

$$\mathop{\mathbf{Pr}}_{X\sim\mathcal{D}_p}(cX > t) \leqslant \mathop{\mathbf{Pr}}_{Y\sim\mathcal{D}}(Y > t).$$

We then define the distribution $\mathcal{D}'_p$ which draws $Z \sim \mathcal{D}'_p$ as cX for $X \sim \mathcal{D}$ if $|cX| > T$, and 0 otherwise. Note then that for $Z \sim \mathcal{D}'_p$ and $Y \sim \mathcal{D}$, $|Y|$ stochastically dominates $|Z|$.

We are then in the position to apply the following theorem from probability theory.

Theorem I.1 (Theorem 2, Pruss (1997)) *Let $X_1, X_2, \ldots, X_d$ be independent symmetric random variables, and suppose $Y_1, Y_2, \ldots, Y_d$ are also independent symmetric random variables. Assume that for every j we have $|Y_j|$ stochastically dominated by $|X_j|$. Then*

$$\mathbf{Pr}\left\{\left|\sum_{j=1}^{d} Y_j\right| \geqslant \lambda\right\} \leqslant 2\,\mathbf{Pr}\left\{\left|\sum_{j=1}^{d} X_j\right| \geqslant \lambda\right\}$$

for every positive λ.

Thus, it suffices to show Equation 4 for $\mathcal{D}'_p$ in place of $\mathcal{D}$. For $j \in [d]$, let $X_j \sim \mathcal{D}_p$ and define

$$\hat{X}_j := \begin{cases} 0 & \text{if } |cX_j| > T \\ X_j & \text{otherwise} \end{cases}.$$

Note then that $X_j - \hat{X}_j \sim \mathcal{D}'_p$, so

$$\mathbf{Pr}\left\{\left|\sum_{j=1}^{d} \mathbf{x}_j Y_j\right| \geqslant \lambda\right\} = \mathbf{Pr}\left\{\left|\sum_{j=1}^{d} \mathbf{x}_j (X_j - X'_j)\right| \geqslant \lambda\right\}.$$

We first have by p-stability that

$$\left|\sum_{j=1}^{d} \mathbf{x}_j X_j\right| \stackrel{d}{=} \|\mathbf{x}\|_p |\hat{X}|$$

for a p-stable variable $\hat{X}$, so there are constants R, p such that

$$\mathbf{Pr}\left(\left|\sum_{j=1}^{d} \mathbf{x}_j X_j\right| \geqslant R\|\mathbf{x}\|_p\right) = \mathbf{Pr}\Big(\|\mathbf{x}\|_p|\hat{X}| \geqslant R\|\mathbf{x}\|_p\Big) = \mathbf{Pr}\Big(|\hat{X}| \geqslant R\Big) \geqslant p.$$

Next note that $X'_j \leqslant T/c = O(1)$ so

$$\mathbf{E}\left|\sum_{j=1}^{d} \mathbf{x}_j X'_j\right| \leqslant \sqrt{\mathbf{E}\left|\sum_{j=1}^{d} \mathbf{x}_j X'_j\right|^2} = \mathbf{E}\sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \mathbf{x}_i\mathbf{x}_j\,\mathbf{E}[X'_iX'_j]} = \sqrt{\sum_{j=1}^{d} \mathbf{x}_j^2\,\mathbf{E}\,X_j'^2} = O(\|\mathbf{x}\|_2)$$

by Jensen's inequality. Then by Markov's inequality, with probability at least $1-p/2$, $\left|\sum_{j=1}^{d} \mathbf{x}_j X'_j\right| \leqslant C\|\mathbf{x}\|_2$ for some constant C that depends on p. Then for $\mathbf{x}$ such that $R\|\mathbf{x}\|_p \geqslant 2C\|\mathbf{x}\|_2$, we have by a union bound that

$$\mathbf{Pr}\left\{\left|\sum_{j=1}^{d} \mathbf{x}_j(X_j - X'_j)\right| \geqslant \frac{R}{2}\|\mathbf{x}\|_p\right\} \geqslant \mathbf{Pr}\left\{\left|\sum_{j=1}^{d} \mathbf{x}_j X_j\right| - \left|\sum_{j=1}^{d} \mathbf{x}_j X'_j\right| \geqslant \frac{R}{2}\|\mathbf{x}\|_p\right\} \geqslant \frac{p}{2}.$$

On the other hand, if $R\|\mathbf{x}\|_p < 2C\|\mathbf{x}\|_2$, the argument in Lemma E.20 shows that

$$\mathbf{Pr}\left(\left|\sum_{j=1}^{d} \mathbf{x}_j Y_j\right| \geqslant \Omega(\|\mathbf{x}\|_2)\right) = \Omega(1)$$

so the result holds under this case as well. ■

Proof [Proof of Theorem E.19] The distortion upper bound is just Lemma E.2.

Mass of small entries. Let $\mathbf{A} = \mathbf{A}^H + \mathbf{A}^L$ as in Definition E.7, with $T = O\big((nd^2\log d/r)^{1/p}\big)$.

By Lemma E.3, the sizes and mass of all level sets $\mathbf{v}_{(k)}$ with entries at most $2^k \leqslant T$ are concentrated around their means up to constant factors with probability at least $1 - \exp(-\Theta(n2^{-kp}))$. Thus by a union bound over d columns j and level sets $0 \leqslant k \leqslant \log_2 T$, with probability at least

$$1 - d\sum_{k=0}^{\log_2 T} \exp\Big(-\Theta(n2^{-kp})\Big) \geqslant 1 - d\exp(-\Theta(\log d)) = 1 - \frac{1}{\mathrm{poly}(r/d^2\log d)}$$

we have for all $j \in [d]$ and $0 \leqslant k \leqslant \log_2 T$ that

$$\|(\mathbf{A}\mathbf{e}_j)_{(k)}\|_0 = \Theta(n2^{-kp}) \qquad \|(\mathbf{A}\mathbf{e}_j)_{(k)}\|_1 = \Theta(n2^{k(1-p)})$$

Then

$$\|\mathbf{A}^L\mathbf{e}_j\|_1 \leqslant \sum_{k=0}^{\log_2 T} \|(\mathbf{A}\mathbf{e}_j)_{(k)}\|_1 = O(n).$$

Mass of large entries. Furthermore, let $\mathbf{B}'$ be the subset of rows of $\mathbf{A}^H$ given by Lemma E.11 that are hashed to locations without any other rows of $\mathbf{A}^H$. Recall also τ_1 and τ_2 from the lemma.

We first have that $\|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 = \Omega(\|\mathbf{A}^H\mathbf{x}\|_1)$ since the rows containing entries larger than τ_1 are perfectly hashed, while rows containing entries between τ_2 and τ_1 are preserved up to constant factors.

Let $\mathbf{B}' = \mathbf{B}'_{>T} + \mathbf{B}'_{\leqslant T}$ where $\mathbf{B}'_{>T}$ contains the entries of $\mathbf{B}'$ that have absolute value greater than T and $\mathbf{B}'_{\leqslant T}$ contains the rest of the entries. Note then that $\mathbf{B}'_{>T}$ has at most one nonzero entry per row, and $\mathbf{B}'_{\leqslant T}$ has at most $O(d \cdot r/d \log d) = O(r/\log d)$ nonzero entries and thus by Lemma E.4, $\|\mathbf{B}'_{\leqslant T}\|_\infty \leqslant O(r^{1/p})$ with probability at least 0.99. We condition on this event. Then for all $\mathbf{x}$,

$$\begin{aligned}
\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 &\geqslant \|\mathbf{S}\mathbf{B}'\mathbf{x}\|_1 \\
&\geqslant \|\mathbf{S}\mathbf{B}'_{>T}\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 \\
&= \sum_{j=1}^{d} |\mathbf{x}_j| \|\mathbf{B}'_{>T}\mathbf{e}_j\|_1 - \|\mathbf{S}\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \mathbf{B}'_{>T}\mathbf{e}_j \text{ have disjoint support} \\
&\geqslant \sum_{j=1}^{d} |\mathbf{x}_j| \sum_{k=\log_2 \tau_2}^{\log_2 \tau_1} 2^k \Theta(n/2^{kp}) - \|\mathbf{B}'_{\leqslant T}\mathbf{x}\|_1 && \text{Lemmas E.11 and E.2} \\
&= \Omega\left(\frac{r}{d^2 \log d}(nd^2 \log d/r)^{1/p}\right)\|\mathbf{x}\|_1 - O(r)\|\mathbf{B}'_{\leqslant T}\|_\infty \|\mathbf{x}\|_1 && \text{Hölder's inequality} \\
&= \Omega\Big((r/d^2 \log d)^{1-1/p} n^{1/p}\Big)\|\mathbf{x}\|_1 - O(r^{1+1/p})\|\mathbf{x}\|_1 \\
&= \Omega((r/d^2 \log d)^{1-1/p} n^{1/p})\|\mathbf{x}\|_1.
\end{aligned}$$

Conclusion. On the other hand, by Lemma E.14, the mass of the $O(r/d \log d)$ rows that are hashed together with the rows of $\mathbf{A}^H$ have mass at most

$$O\left(\frac{1}{\sqrt{\log d}} \frac{n^{1/p}}{(r/d^2 \log d)^{1/p-1}}\right)\|\mathbf{x}\|_1 = o\Big((r/d^2 \log d)^{1-1/p} n^{1/p}\Big)\|\mathbf{x}\|_1.$$

Then,

$$\begin{aligned}
\frac{1}{\kappa} &\geqslant \frac{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{A}\mathbf{x}\|_1} \geqslant \frac{\|\mathbf{S}\mathbf{A}^H\mathbf{x}\|_1 - \|\mathbf{S}\mathbf{C}_1\mathbf{x}\|_1}{\|\mathbf{A}^H\mathbf{x}\|_1 + \|\mathbf{A}^L\mathbf{x}\|_1} \\
&\geqslant \frac{\Omega(\|\mathbf{A}^H\mathbf{x}\|_1 + (r/d^2 \log d)^{1-1/p} n^{1/p})\|\mathbf{x}\|_1}{O(\|\mathbf{A}^H\mathbf{x}\|_1 + n)\|\mathbf{x}\|_1} \geqslant \Omega\left(\left(\frac{(r/d^2 \log d)}{n}\right)^{1-1/p}\right).
\end{aligned}$$

■

I.1. Proofs for Section E.5

Proof [Proof of Lemma E.20] For the upper bound, we have by Jensen's inequality that

$$\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}^d} |\langle \mathbf{v}, \mathbf{x}\rangle| \leqslant \sqrt{\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}^d} |\langle \mathbf{v}, \mathbf{x}\rangle|^2} = \sqrt{\sum_{j=1}^{d} \mathbf{x}_j^2 \mathbf{E}\, \mathbf{v}_j^2} = O(\|\mathbf{x}\|_2).$$

We now focus on the lower bound.

Let $M = O(1)$ be the median of $\mathcal{D}$. We define $\mathbf{w}$ to be the truncation of $\mathbf{v}$ at M, that is, $\mathbf{w}_i = 0$ if $|\mathbf{v}_i| > M$ and $\mathbf{w}_i = \mathbf{v}_i$ otherwise. Then by (Vershynin, 2018, Lemma 6.1.2),

$$\mathbf{E}|\langle \mathbf{v}, \mathbf{x}\rangle| \geqslant \mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|$$

so it suffices to bound $\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|$ instead.

Note that

$$\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} E(\mathbf{w}_i\mathbf{x}_i\mathbf{w}_j\mathbf{x}_j) = \sum_{j=1}^{d} \mathbf{x}_j^2 \,\mathbf{E}\,\mathbf{w}_j^2 = \Omega(\|\mathbf{x}\|_2^2)$$

and

$$\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^4 = \sum_{j=1}^{d} \mathbf{x}_j^4 \,\mathbf{E}\,\mathbf{w}_j^4 + 3\sum_{j\neq k}^{d} \mathbf{x}_j^2\mathbf{x}_k^2\,\mathbf{E}(\mathbf{w}_j^2\mathbf{w}_k^2) \leqslant O(\|\mathbf{x}\|_4^4 + \|\mathbf{x}\|_2^4) = O(\|\mathbf{x}\|_2^4)$$

so by the Paley-Zygmund inequality,

$$\mathbf{Pr}\Big(|\langle \mathbf{w}, \mathbf{x}\rangle| \geqslant \sqrt{\lambda}\sqrt{\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^2}\Big) = \mathbf{Pr}\big(|\langle \mathbf{w}, \mathbf{x}\rangle|^2 \geqslant \lambda\,\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^2\big) \geqslant (1-\lambda)^2\frac{(\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^2)^2}{\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle|^4} = \Omega(1).$$

Thus $|\langle \mathbf{w}, \mathbf{x}\rangle| = \Omega(\|\mathbf{x}\|_2)$ with constant probability and thus $\mathbf{E}|\langle \mathbf{w}, \mathbf{x}\rangle| = \Omega(\|\mathbf{x}\|_2)$, as desired. ■

Proof [Proof of Lemma E.21] Let $X \coloneqq \langle \mathbf{v}, \mathbf{x}\rangle$. We have

$$\mathbf{Pr}(\neg\mathcal{E}_i) = B^{-p} \leqslant \frac{\varepsilon}{d}$$

so by the union bound,

$$\mathbf{Pr}(\mathcal{E}) \geqslant 1 - \sum_{i=1}^{d} \mathbf{Pr}(\neg\mathcal{E}_i) = 1 - d\,\mathbf{Pr}(\neg\mathcal{E}_1) = 1 - \varepsilon.$$

For B large enough, we have by the layer cake theorem that

$$\mathop{\mathbf{E}}_{Y\sim\mathcal{D}}(|Y| \mid |Y| > B) \leqslant \frac{1}{\mathbf{Pr}(|Y| > B)}\int_B^\infty O(x^p)\,dx = \frac{1}{\Omega(B^{-p})}O(B^{1-p}) = O(B)$$

since $p \geqslant 2 > 1$. Then,

$$\begin{aligned}
\mathbf{E}(|X| \mid \neg\mathcal{E}) &\leqslant \sum_{i=1}^{d} \mathbf{E}(|\mathbf{v}_i\mathbf{x}_i| \mid \neg\mathcal{E}) \\
&= \sum_{i=1}^{d} \mathbf{E}(|\mathbf{v}_i\mathbf{x}_i| \mid \neg\mathcal{E}_i, \neg\mathcal{E})\,\mathbf{Pr}(\neg\mathcal{E}_i \mid \neg\mathcal{E}) + \mathbf{E}(|\mathbf{v}_i\mathbf{x}_i| \mid \mathcal{E}_i, \neg\mathcal{E})\,\mathbf{Pr}(\mathcal{E}_i \mid \neg\mathcal{E}) \\
&\leqslant \sum_{i=1}^{d} O(B)|\mathbf{x}_i|\frac{1}{d} + O(|\mathbf{x}_i|) \\
&\leqslant O(B+d)\|\mathbf{x}\|_1.
\end{aligned}$$

We then have

$$\mathbf{E}|X| = \mathbf{E}(|X| \mid \mathcal{E})\,\mathbf{Pr}(\mathcal{E}) + \mathbf{E}(|X| \mid \neg\mathcal{E})\,\mathbf{Pr}(\neg\mathcal{E}) \leqslant \mathbf{E}(|X| \mid \mathcal{E}) + O(B+d)\|\mathbf{x}\|_1\,\mathbf{Pr}(\neg\mathcal{E}).$$

Since $\mathbf{E}|X| = \Omega(\|\mathbf{x}\|_2) = \Omega(\|\mathbf{x}\|_1/\sqrt{d})$ by Lemma E.20,

$$O(B+d)\|\mathbf{x}\|_1\,\mathbf{Pr}(\neg\mathcal{E}) = O(B+d)\|\mathbf{x}\|_1\frac{1}{B^p} \leqslant O(\varepsilon)\,\mathbf{E}|X|$$

by our choice of B. We thus have

$$\mathbf{E}|X| \leqslant \mathbf{E}(|X| \mid \mathcal{E}) + O(\varepsilon)\,\mathbf{E}|X|$$

so

$$\mathbf{E}(|X| \mid \mathcal{E}) \geqslant (1 - O(\varepsilon))\,\mathbf{E}|X|.$$

■