On Sketching the q to p Norms

Aditya Krishnan
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA
arkrishn@andrew.cmu.edu

Sidhanth Mohanty
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA
sidhanth@cmu.edu

David P. Woodruff
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA
dwoodruf@cs.cmu.edu

Abstract

We initiate the study of data dimensionality reduction, or sketching, for the $q \to p$ norms. Given an $n \times d$ matrix A, the $q \to p$ norm, denoted $\|A\|_{q\to p} = \sup_{x\in\mathbb{R}^d\setminus\vec{0}} \frac{\|Ax\|_p}{\|x\|_q}$, is a natural generalization of several matrix and vector norms studied in the data stream and sketching models, with applications to datamining, hardness of approximation, and oblivious routing. We say a distribution S on random matrices $L \in \mathbb{R}^{nd} \to \mathbb{R}^k$ is a (k, α)-sketching family if from $L(A)$, one can approximate $\|A\|_{q\to p}$ up to a factor α with constant probability. We provide upper and lower bounds on the sketching dimension k for every $p, q \in [1, \infty]$, and in a number of cases our bounds are tight. While we mostly focus on constant α, we also consider large approximation factors α, as well as other variants of the problem such as when A has low rank.

2012 ACM Subject Classification Theory of computation → Numeric approximation algorithms

Keywords and phrases Dimensionality Reduction, Norms, Sketching, Streaming

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2018.15

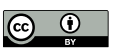
Related Version A full version of the paper is available at `https://arxiv.org/abs/1806.06429`.

Funding D. Woodruff would like to acknowledge the support by the National Science Foundation under Grant No. CCF-1815840.

1 Introduction

Data dimensionality reduction, or sketching, is a powerful technique by which one compresses a large dimensional object to a much smaller representation, while preserving important structural information. Motivated by applications in streaming and numerical linear algebra, the object is often a vector $x \in \mathbb{R}^n$ or a matrix $A \in \mathbb{R}^{n\times d}$. One of the most common forms of sketching is oblivious sketching, whereby one chooses a random matrix L from some distribution S, and compresses x to Lx or A to $L(A)$. The latter quantity $L(A)$ denotes a linear map from $\mathbb{R}^{nd}$, interpreting A as an nd-dimensional vector, to an often much lower dimensional space, say $\mathbb{R}^k$ for a value $k \ll nd$.

Sketching has numerous applications. For example, in the data stream model, one sees additive updates $x_i \leftarrow x_i + \Delta$, where the update indicates that x_i should change from its old value by an additive Δ. Given a sketch $L \cdot x$, one can update it by replacing it with $L \cdot x + \Delta \cdot L_{*,i}$, where $L_{*,i}$ denotes the i-th column of L. Thus, it is easy to maintain a sketch of a vector evolving in the streaming model. Similarly, in the matrix setting, given an



update $A_{i,j} \leftarrow A_{i,j} + \Delta$, one can update $L(A)$ to $L(A) + \Delta L(e_{i,j})$, where $e_{i,j}$ denotes the matrix with a single one in the (i,j)-th position, and is otherwise 0. If L is oblivious, that is, sampled from a distribution independent of x (or A in the matrix case), then one can create L without having to see the entire stream in advance. Other applications include distributed computing, whereby a vector or matrix is partitioned across multiple servers. For instance, server 1 might have a vector x^1 and server 2 a vector x^2. Given the sketches Lx^1 and Lx^2, by linearity one can combine them, using $L(x^1 + x^2) = Lx^1 + Lx^2$. In these applications it is important that the number k of rows of L is small, since it is proporational to the memory required of the data stream algorithm, or the communication in a distributed protocol. Here k is referred to as the *sketching* dimension.

Sketching vector norms is fairly well understood, and we have tight bounds up to logarithmic factors for estimating the ℓ_p-norms $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ for every $p \in [1, \infty]$; for a sample of such work, see [1, 10, 24, 23, 28, 27] for work in the related data stream context, and [40, 9, 33] for work specifically in the sketching model. Recently, there is work [13] characterizing the sketching complexity of any symmetric norm on a vector x. A number of works have also looked at sketching *matrix norms*. In particular, the Schatten p-norms $\|A\|_p = \left(\sum_{i=1}^{\text{rank}(A)} \sigma_i(A)^p\right)^{1/p}$ have gained considerable attention. They have proven to be considerably harder to approximate than the vector p-norms, and understanding their complexity has led to important algorithmic and lower bound techniques. A body of work has focused on understanding the complexity of estimating matrix norms in the data stream model with 1-pass over the stream [4, 34], as well as with multiple passes [15], the sketching model [32, 36], statistical models [31, 29], as well as the general RAM model [38, 44]. Dimensionality reduction in these norms also has applications in quantum computing [46, 22], and are studied in nearest neighbor search data structures [2].

1.1 Our Contributions

We consider the sketching complexity of a new family of norms, namely, the $p \to q$ norms of a matrix. A common quantity that arises in various applications is the amount by which a linear map A "stretches" vectors. One way to measure this quantity is the maximum singular value of A, which can be written as $\sup_{\|x\|_2=1} \|Ax\|_2$, and is just the Schatten-∞ norm, defined above. In this work we consider a different way of measuring this stretch, which considerably generalizes the operator norm.

For a linear operator A from a normed space $\mathcal{X}$ to a normed space $\mathcal{Y}$, we define $\|A\|_{\mathcal{X}\to\mathcal{Y}}$ as $\sup_{\|x\|_{\mathcal{X}}=1} \|Ax\|_{\mathcal{Y}}$. Of specific interest to us is the case where $\mathcal{X} = \ell_q^d$ and $\mathcal{Y} = \ell_p^n$, and we denote the corresponding norm of such an operator by $\|A\|_{q\to p}$. Our objective is to study the sketching complexity of approximating this norm.

▶ **Definition 1** ((k,α)-sketching family)**.** Let $\mathcal{S}$ be a distribution over linear functions from $\mathbb{R}^{n\times d}$ to $\mathbb{R}^k$ and f a function from $\mathbb{R}^k$ to $\mathbb{R}$. We call $(\mathcal{S}, f)$ a **(k,α)-sketching family** for the $q \to p$ norm if for all $A \in \mathbb{R}^{n\times d}$, $\mathbf{Pr}_{L\sim\mathcal{S}}\left[f(L(A)) \in (1/\alpha, \alpha)\,\|A\|_{q\to p}\right] \geq \frac{5}{6}$.

We provide upper and lower bounds on k. The details of the specific results we have are described in Section 1.3.

1.2 Motivation

This problem is well-studied in mathematics when $p = q$ as it simply corresponds to p-matrix norm estimation[1].

An intriguing question is whether one can preserve $\|Ax\|_p$ in a lower-dimensional sketch space, given that the vectors x come from the unit ball of a smaller norm.

Apart from being mathematically interesting, this problem has a number of applications. The operator norm is a special case when $p = q = 2$. The operator norm can be accurately estimated by any subspace embedding for ℓ_2, discussed in detail in [18]. The dual of this norm is also the Schatten-1 norm, which has received considerable attention in the streaming model [34, 15]. The $q \to p$ norm problem is a natural generalization of the operator norm problem, and when $p < 2$, may be more appropriate in the context of robust statistics, where it is known that the p norm for $p < 2$ is less sensitive to outliers, see, e.g., Chapter 3 of [47] for a survey on robust regression, and [42] for recent work on ℓ_1-low rank approximation.

The $2 \to q$ norms arise in the hardness of approximation literature and an algorithm for some instances of the problem was used to break the Khot-Vishnoi Unique Games candidate hard instance [30]. Work by [11] gives an algorithm running in time $\exp(n^{2/p})$ for approximating $2 \to p$ norms for all $p \geq 4$. These algorithms give a constant factor approximation when promised the $2 \to p$ norm is in a certain range (depending on the operator norm) rather than providing a general estimate of the $2 \to p$ norm. This same paper also discusses assumptions on the the NP-hardness and ETH hardness of approximating $2 \to p$ norms. The work of [14] extends that of [11] to all $p \geq 2$. The work of [12] gives a PTAS for computing $\|A\|_{q \to p}$ if $1 \leq p \leq q$ and A has non-negative entries, and gives an application of this to the oblivious routing problem where congestion is measured using the ℓ_p norm. The paper also shows that it is hard to approximate $\|A\|_{q \to p}$ within a constant factor for general A, and general p and q. Sketching may allow, for example, for reducing the original problem to a smaller instance of the same problem, which although may still involve exhaustive search, could give a faster concrete running time.

The $1 \to q$ norm turns out to be the maximum of the q-norm of the columns of A, which is related to the heavy hitters problems in data streams, e.g., the column with the largest q-norm may be the most significant or desirable in an application. Likewise, the $q \to \infty$ norms turn out to be the maximum of the p-norms of the rows of A, where p is the dual norm to q, and therefore have similar heavy hitter applications. The $\infty \to q$ norm is maximized when $x \in \{-1, 1\}^n$ and therefore includes the cut-norm as a special case, and is related to Grothendieck inequalities, see, e.g., [16, 39, 17].

Our main motivation for studying the $p \to q$ norms comes from understanding and developing new techniques for this family of norms. Another family of norms that is well-studied in the data stream literature are the *cascaded norms*, which for an $n \times d$ matrix A and parameters p and q, are defined to be $(\sum_{i=1,\ldots,n}(\|A_{i,*}\|_p)^q)^{1/q}$, where $A_{i,*}$ denotes the i-th row of A. That is, we compute the q-norm of the vector of p-norms of the rows of A. This problem originated in [19] and has applications to mining multi-graphs; the following sequence of work established tight bounds up to logarithmic factors for every $p, q \in [1, \infty]$ [26, 6]. This line of work led to very new techniques; one highlight is the use of Poincaré inequalities in proving information complexity lower bounds, which has then been studied in a number of followup works [5, 25, 7].

[1] See, e.g., `https://en.wikipedia.org/wiki/Matrix_norm`

1.3 Our Results

After establishing preliminary results and theorems in Section 2, we give our results for constant and large approximation factors. Our main theorem is as follows. Here ℓ_{q^*} is the dual norm of ℓ_q, that is, $1/q^* + 1/q = 1$ (when $q = 1$, $q^* = \infty$, and vice versa).

▶ **Theorem 2.** *For all matrices $A \in \mathbb{R}^{n\times n}$ with rank r and real values $p, q \in [1, \infty]$, the table below gives upper and lower bounds on k for a $(k, \Theta(1))$-sketching family of various $q \to p$ norms.*

$q \to p$ Norm	$p^* \to q^*$ Norm	Upper Bound	Sec	Lower Bound	Sec
$1 \to [1,2]$	$[2,\infty] \to \infty$	$O(n \log n)$	*3.1*	$\Omega(n)$	*4.2*
$1 \to [2,\infty]$	$[1,2] \to \infty$	$O(n^{2-\frac{2}{p}} \log^2 n)$	*3.1*	$\Omega(n^{2-\frac{2}{p}})$	*4.3*
$[2,\infty] \to [1,2]$	$[2,\infty] \to [1,2]$	$O(n^2)$	-	$\Omega(n^2)$	*4.4*
$2 \to [2,\infty]$	$[1,2] \to 2$	$O(\min\{n^{1-\frac{2}{p}} r^2 \log n, n^2\})$	*3.2*	$\Omega(\min\{n, n^{1-\frac{2}{p}} r\})$	*4.5*
$[1,2] \to [1,2]$	$[2,\infty] \to [2,\infty]$	$O(n^2)$	-	$\Omega(\min\{n^{1-\frac{2}{q^*}} r, n\})$	*4.5*
$[1,2] \to [2,\infty]$	$[1,2] \to [2,\infty]$	$O(n^2)$	-	$\Omega\left(\frac{n}{\log n}\right)$	*4.6*

The constant factor hidden in Theorem 2 does not hold for all constants, the smallest constant it holds for varies depending on the specific values of q, p.

We also have several results for large approximation factors summarized in the theorem below.

▶ **Theorem 3.** *There exists a $\left(O\left(\frac{n^2}{\alpha}\right), \alpha\right)$-sketching family for the $2 \to p$ and $\infty \to p$ norm and a $\left(O\left(\frac{n^2}{\alpha^2}\right), \alpha\right)$-sketching family for the $q \to p$ norm for $q \geq 1$ and $1 \leq p \leq 2$.*

Our algorithms combine several insights, which we illustrate here in the case of the $2 \to p$ norm for $p \geq 2$ and when the rank of A is r: (1) we show by duality that $\|A\|_{2\to p}$ is the same as $\|A^T\|_{p^*\to 2}$, where p^* satisfies $\frac{1}{p^*} + \frac{1}{p} = 1$ and is the dual norm to p. Although the proof is elementary, this plays several key roles in our argument. Next, we (2) use oblivious subspace embeddings S which provide constant factor approximations for all vectors simultaneously in an r-dimensional subspace of ℓ_2, and enable us to say that with Cr rows for a constant $C > 0$, we have $\|SA^T\|_{p^*\to 2} = \Theta(1)\|A^T\|_{p^*\to 2}$. Next, (3) we use that for a random Gaussian matrix $G \in \mathbb{R}^{C'r\times Cr}$, for a constant $C' > 0$, with appropriate variance, it has the property that simultaneously for all $x \in \mathbb{R}^{Cr}$, $\|Gx\|_1 = \Theta(1) \cdot \|x\|_2$. This is a special case of Dvoretsky's theorem in functional analysis. Thus, instead of directly approximating $\|SA^T\|_{p^*\to 2}$, we can obtain a constant factor approximation by approximating $\|GSA^T\|_{p^*\to 1}$. This is another norm we do not know how to directly work with, so we apply duality (1) again, and argue this is the same as approximating $\|AS^TG^T\|_{\infty\to p}$. A key observation is now (4), that $\sup_{x \text{ s.t. } \|x\|_\infty = 1} \|AS^TG^Tx\|_p$ is realized when x has each coordinate equal to 1 or -1. Consequently, as $x \in \mathbb{R}^{C'r}$, it suffices to use any sketch T for the p-norm of a fixed vector which fails with probability $\exp(-C'r)$, and estimate $\|TAS^TG^Tx\|_p$ for each of the $2^{C'r}$ possible maximizers x, and output the largest estimate. As there exist sketches T with $O(n^{1-2/p}r \log n)$ rows for this purpose, this gives us an overall sketching complexity of $O(n^{1-2/p}r^2 \log n)$.

We defer a discussion of our lower bound techniques to Section 4.

2 Preliminaries

In this section, we introduce the tools we use in this paper.

▶ **Definition 4** (Total Variation Distance)**.** Given two distributions $\mathcal{D}$ and $\mathcal{D}'$ over sample space Ω with density functions $p_{\mathcal{D}}$ and $p_{\mathcal{D}'}$, the **total variation distance** is defined in two equivalent ways as follows $d_{TV}(\mathcal{D}, \mathcal{D}') = \frac{1}{2}\|p_{\mathcal{D}} - p_{\mathcal{D}'}\|_1 = \sup_{\mathcal{E}} |\mathbf{Pr}_{x\sim\mathcal{D}}[\mathcal{E}] - \mathbf{Pr}_{x\sim\mathcal{D}'}[\mathcal{E}]|$

The following result bounds the total variation distance between two multivariate Gaussians.

▶ **Lemma 5** ([21], Lemma A4)**.** *Let λ be the minimum eigenvalue of PSD matrix Σ, then $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')) \leq \frac{C}{\sqrt{\lambda}}(\|\mu - \mu'\|_2 + \|\Sigma - \Sigma'\|_F)$ for an absolute constant C.*

We state a well known result that a Lipschitz function of a Gaussian vector is tightly concentrated around its expectation, which is useful since ℓ_p norms are Lipschitz.

▶ **Theorem 6** ([43], Theorem 2.1.12)**.** *Let $X \sim \mathcal{N}(0, I_n)$ be a Gaussian random vector and let $f : \mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz function. Then for some absolute constants $C, c > 0$, $\mathbf{Pr}[|f(X) - \mathbf{E}[f(X)]| \geq \lambda] \leq C\exp(-c\lambda^2)$ Notice that this implies if f is t-Lipschitz, then $\mathbf{Pr}[|f(X) - \mathbf{E}[f(X)]| \geq \lambda] \leq C\exp(-c\lambda^2/t^2)$*

It is possible to embed ℓ_2^n into $\ell_p^{O(n)}$ with constant distortion using a linear map when $p \in [1, 2]$, and we use the existence of such a linear map in our results.

▶ **Lemma 7** ([37], Theorem 2.5.1)**.** *For all $p \in [1, 2]$, there is an absolute constant C_p such that for any n, there is a linear map $T : \mathbb{R}^n \to \mathbb{R}^{C_p n}$ such that $\|T(x)\|_p = \left(1 \pm \frac{1}{2}\right)\|x\|_2$. An important observation is that this implies for any linear map $A : \mathbb{R}^n \to \mathbb{R}^n$, we have $\|TA\|_{q\to p} = \left(1 \pm \frac{1}{2}\right)\|A\|_{q\to 2}$.*

In the lemma below we make an important observation that highlights the connection between several $p \to q$ norms.

▶ **Lemma 8.** *For any $p, q \geq 1$ and $d \times n$ matrix A, $\|A\|_{q\to p} = \|A^T\|_{p^*\to q^*}$.*

Proof. Using the notation above for dual norms, we have

$$\begin{aligned}
\|A\|_{q\to p} &= \sup\{\|Ax\|_q : \|x\|_p \leq 1\} \\
&= \sup\{\sup\{y^\top Ax : \|y\|_{q*} \leq 1\} : \|x\|_p \leq 1\} \\
&= \sup\{\sup\{x^\top A^\top y : \|x\|_p \leq 1\} : \|y\|_{q^*} \leq 1\} \\
&= \sup\{\|A^\top y\|_{p^*} : \|y\|_{q^*} \leq 1\} \\
&= \|A^\top\|_{p^*\to q^*}
\end{aligned}$$

◀

Throughout the paper, we make use of q^* to refer to $\frac{q}{q-1}$ since $\ell_{\frac{q}{q-1}}$ is the dual norm of ℓ_q.

We give a characterization of the $1 \to p$ and $\infty \to p$ norm of a matrix. The proofs can be found in the full version's Appendix A. For any $d \times n$ matrix A, we have

▶ **Lemma 9.** $\|A\|_{1\to p} = \max_{i\in[n]}\{\|A_{*,i}\|_p\}$.

▶ **Lemma 10.** $\|A\|_{\infty\to p} = \max_{x\in\{\pm 1\}^n}\|Ax\|_p$.

We introduce the machinery of ε-nets, a common tool in the study of random matrices (see [45]) along with some relevant lemmas and defer the proofs to the full version's Appendix.

▶ **Definition 11** (ε-net)**.** Let $\mathcal{X}$ be a normed space. For $S \subseteq V$, we call a set N an ε-net for S if for all $v \in S$, there is $v' \in N$ such that $\|v - v'\|_{\mathcal{X}} < \varepsilon$.

For a linear operator A, we show that to bound $\|A\|_{\mathcal{X}\to\mathcal{Y}}$, it suffices to bound $\|Ax\|_{\mathcal{Y}}$ for x taken over an ε-net of the unit ball in $\mathcal{X}$.

▶ **Lemma 12.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces and let $A : \mathcal{X} \to \mathcal{Y}$ be a linear map. Suppose N is an ε-net of the unit ball in $\mathcal{X}$, then $\|A\|_{\mathcal{X}\to\mathcal{Y}} \leq \frac{1}{1-\varepsilon} \max_{v\in N} \|Av\|_{\mathcal{Y}}$.*

We also give a way to construct 'small' ε-nets of unit balls.

▶ **Lemma 13.** *There is an ε-net of the unit ball B in an n-dimensional normed space $\mathcal{X}$ with at most $\left(\frac{2+\varepsilon}{\varepsilon}\right)^n$ elements.*

Another tool we use is subspace embeddings, which we define below.

▶ **Definition 14.** An **oblivious subspace embedding family** (OSE family) is a distribution $\mathcal{S}$ over $O(m) \times n$ matrices such that for any subspace $K \subseteq \mathbb{R}^n$ of dimension m, $\mathbf{Pr}_{S\sim\mathcal{S}}[\forall x \in K : \|Sx\|_2 = \Theta(1)\|x\|_2] \geq \frac{9}{10}$.

▶ **Lemma 15** ([41])**.** *There exist OSE families, where the matrices have dimension $O(k) \times n$. Note that this means for any rank-k matrix A, a randomly drawn S from such an oblivious subspace embedding family satisfies $\|SAx\|_2 = \Theta(1)\|Ax\|_2$ simultaneously for all x with probability at least $99/100$.*

3 Sketching algorithms for constant factor approximations

3.1 Sketches for approximating $\|A\|_{1\to p}$

We show how to use sketches for p-norms of vectors to come up with sketches for the $1 \to p$ norm.

▶ **Lemma 16.** *Let x be an arbitrary vector in $\mathbb{R}^n$. If $\mathcal{S}$ is a distribution over $t \times n$ sketching matrices, and $f : \mathbb{R}^t \to \mathbb{R}$ is a function such that $\mathbf{Pr}_{S\sim\mathcal{S}}\left[f(Sx) \in \left(\frac{1}{2}\|x\|_p, 2\|x\|_p\right)\right] \geq \frac{2}{3}$ then there is an $(O(nt\log n), 2)$-sketching family $(\mathcal{S}', g)$ for the $1 \to p$ norm of $n \times n$ matrices.*

Proof. Proof in the full version's Appendix B. ◀

Given an n-dimensional vector x, we have the following theorems from [28] and [6] respectively.

▶ **Theorem 17** (Efficient sketches for small norms)**.** *When $p \in [1, 2]$, there is a function f and a distribution over sketching matrices $\mathcal{F}$ with $O(1)$ rows such that for $S \sim \mathcal{F}$, $f(Sx)$ is a constant factor approximation for $\|x\|_p$ with probability at least $2/3$.*

▶ **Theorem 18** (Efficient sketches for large norms)**.** *When $p > 2$, there is a function f and a distribution over sketching matrices $\mathcal{F}$ with $O(n^{1-2/p}\log n)$ rows such that for $S \sim \mathcal{F}$, $f(Sx)$ is a constant factor approximation for $\|x\|_p$ with probability at least $2/3$.*

Lemma 16 tells us the following as a corollary to Theorems 17 and 18.

▶ **Theorem 19.** *There is an $(O(n\log n), 2)$-sketching family for the $1 \to p$ norm when $p \in [1, 2]$ and a $(O(n^{2-2/p})\log^2 n, 2)$-sketching family for the $1 \to p$ norm when $p \in (2, \infty]$.*

3.2 Sketches for approximating $\|A\|_{2\to p}$ for $p > 2$

We give a sketching algorithm for the $2 \to p$ norm of A, whose number of measurements depends on the rank r of $d \times n$ matrix A.

▶ **Theorem 20.** *There is an $(O(n^{1-2/p}r^2\log n), \Theta(1))$-sketching family for the $2 \to p$ norm.*

Proof. Observe that $\|A\|_{2\to p}$ is equal to $\|A^T\|_{p^*\to 2}$ by Lemma 8 and let S be a $Cr \times d$ matrix drawn from an oblivious subspace embedding family, which exists by Lemma 15. From Lemma 7, let G be a $\beta r \times Cr$ map such that for all x, $\|GSA^Tx\|_1 = \Theta(1)\|SA^Tx\|_2$. Combining with the subspace embedding property, we get that $\|GSA^Tx\|_1 = \Theta(1)\|A^Tx\|_2$ for all x, which is equivalent to saying $\|GSA^T\|_{p^*\to 1} = \Theta(1)\|A\|_{2\to p}$. Another application of Lemma 8 gives us that $\|AS^TG^T\|_{\infty\to p} = \Theta(1)\|A\|_{2\to p}$. Since AS^TG^T is $n \times \beta r$, $\|AS^TG^T\|_{\infty\to p} = \max_{x\in\{\pm 1\}^{\beta r}} \|AS^TG^Tx\|_p$.

Our final ingredient is the existence of an $O(n^{1-2/p}\log n \log(1/\delta)) \times n$ sketching matrix E and estimation function f such that for any x, $\mathbf{Pr}[f(Ey) = \Theta(1)\|y\|_p] \geq 1-\delta$ [3] when $p > 2$. We set $\delta = 2^{-2\beta r}$ and use a union bound over all $2^{\beta r}$ vectors in $\{\pm 1\}^{\beta r}$ to conclude

$$\mathbf{Pr}[\forall x \in \{\pm 1\}^{\beta r} : f(EAS^TG^Tx) = \Theta(1)\|AS^TG^Tx\|_q] \geq 1 - 2^{-\beta r}$$

$$\mathbf{Pr}\left[\max_{x\in\{\pm 1\}^{\beta r}} f(EAS^TG^Tx) = \Theta(1)\|AS^TG^T\|_{\infty\to q}\right] \geq 1 - 2^{-\beta r}$$

Consequently, we get a sketch that consists of $O(n^{1-2/p}r^2\log n)$ measurements to get a $\Theta(1)$ approximation to $\|A\|_{2\to p}$ with probability at least 0.99. ◀

4 Sketching lower bounds for constant factor approximations

4.1 Lower Bound Techniques

The way we prove most of our lower bounds is by giving two distributions over $n \times n$ matrices, $\mathcal{D}_1$ and $\mathcal{D}_2$, where matrices drawn from the two distributions have $q \to p$ norm separated by a constant factor κ with high probability, which means a $(k, \sqrt{\kappa})$-sketching family can distinguish between samples from the two distributions. We then show an upper bound on the variation distance between distributions of k-dimensional sketches of $\mathcal{D}_1$ and $\mathcal{D}_2$. We then argue that if k is too small, then the total variation distance is too small to solve the distinguishing problem. We formalize this intuition in the following theorem.

▶ **Theorem 21.** *Suppose $\mathcal{D}_1$ and $\mathcal{D}_2$ are distributions over $d \times n$ matrices such that*

(i) $\mathbf{Pr}_{D\sim\mathcal{D}_1}[\|D\|_{q\to p} < s] \geq 1 - \frac{1}{n}$ *and* $\mathbf{Pr}_{D\sim\mathcal{D}_2}[\|D\|_{q\to p} > \kappa s] \geq 1 - \frac{1}{n}$

(ii) *for any linear map $L : \mathbb{R}^{d\times n} \to \mathbb{R}^k$, $d_{TV}(L(\mathcal{D}_1), L(\mathcal{D}_2)) = O\left(\frac{k^a}{n^b}\right)$*

for constants s, κ, a, b, any $(k, \sqrt{\kappa})$-sketching family for the $q \to p$ norm must satisfy $k = \Omega(n^{b/a})$.

Proof. Let $\mathcal{D}$ be the distribution over matrices given by sampling from $\mathcal{D}_1$ with probability $\frac{1}{2}$ and drawing from $\mathcal{D}_2$ with probability $\frac{1}{2}$. We shall fix a sketching operator $L : \mathbb{R}^{d\times n} \to \mathbb{R}^k$ and consider A drawn from a distribution $\mathcal{D}$. Suppose $f(L(A))$ lies in $(1/\sqrt{\kappa}, \sqrt{\kappa})\|A\|_{q\to p}$ with probability at least $5/6$. It suffices to show that k must be $\Omega(n^{b/a})$ since the theorem statement then follows from Yao's minimax principle. We must have

$$\mathbf{Pr}_{A\sim\mathcal{D}_1}\left[f(L(A)) \in \left(\frac{1}{\sqrt{\kappa}}, \sqrt{\kappa}\right)\|A\|_{q\to p}\right] \geq \frac{2}{3},$$

$$\mathbf{Pr}_{A\sim\mathcal{D}_2}\left[f(L(A)) \in \left(\frac{1}{\sqrt{\kappa}}, \sqrt{\kappa}\right)\|A\|_{q\to p}\right] \geq \frac{2}{3}$$

Thus, we have an algorithm that correctly distinguishes with probability at least $\frac{3}{5}$ if A was drawn from $\mathcal{D}_1$ or $\mathcal{D}_2$ by checking if $f(L(A))$ is greater than or less than $\sqrt{\kappa}s$.

The existence of this distinguishing algorithm means the total variation distance between the distributions of $L(D_1)$ and $L(D_2)$ is at least $\frac{1}{5}$. From the theorem's hypothesis, we know of a constant C such that $\frac{Ck^a}{n^b} \geq \frac{1}{5}$, which gives us the desired upper bound. ◀

We also show an upper bound on the variation distance of sketches for two distributions that we use throughout this paper. Define $\mathcal{G}_{1,d\times n}$ as the distribution over $d \times n$ Gaussian matrices and $\mathcal{G}_{2,d\times n}[\alpha]$ as the distribution given by drawing a Gaussian matrix and adding αu, where u is a d-dimensional Gaussian vector to a random column. We write $\mathcal{G}_i$ instead of $\mathcal{G}_{i,d\times n}$ when the dimensions of the random matrix are evident from context.

▶ **Lemma 22.** *Let L be a linear sketch from $\mathbb{R}^{d\times n} \to \mathbb{R}^k$ and let $\mathcal{H}_i$ be the distribution of $L(x)$ where x is drawn from $\mathcal{G}_i$. Then $d_{TV}(\mathcal{H}_1, \mathcal{H}_2) \leq \frac{C\alpha^2 k}{n}$ for an absolute constant C.*

Proof. We can think of L as a $k \times nd$ matrix that acts on a sample from $\mathcal{G}_1$ or $\mathcal{G}_2$ as though it were an nd-dimensional vector. Without loss of generality, we can assume that the rows of L are orthonormal, since one can always perform a change of basis in post-processing. Thus, the distribution $\mathcal{H}_1$ is the same as $\mathcal{N}(0, I_k)$. For fixed i and G a $d \times n$ matrix of unit Gaussians, the distribution of $L(G + \alpha u e_i^T)$ is Gaussian with covariance $\mathbf{E}[L(G + \alpha u e_i^T)L(G + \alpha u e_i^T)^T]$, equal to $I + \alpha^2 L_{B_i} L_{B_i}^T$ where L_{B_i} is the submatrix given by columns of L indexed $(i-1)d+1, (i-1)d+2, \ldots, id$. Let $\mathcal{H}_{2,i}$ be $\mathcal{N}(0, I + \alpha^2 L_{B_i} L_{B_i}^T)$. $\mathcal{H}_2$ is the distribution of picking a random i and drawing a matrix from $\mathcal{N}(0, I + L_{B_i} L_{B_i}^T)$.

We now analyze the total variation distance between $\mathcal{H}_1$ and $\mathcal{H}_2$ and get the desired bound from a chain of inequalities. $d_{TV}(\mathcal{H}_1, \mathcal{H}_2) = \frac{1}{2}\int_{x\in\mathbb{R}^k} |p_{\mathcal{H}_1}(x) - p_{\mathcal{H}_2}(x)|dx$

$\leq \frac{1}{2}\int_{x\in\mathbb{R}^k} \left|\sum_{i=1}^n \frac{1}{n} p_{\mathcal{H}_1}(x) - \frac{1}{n} p_{\mathcal{H}_{2,i}}(x)\right| dx \leq \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\int_{x\in\mathbb{R}^k} \left|p_{\mathcal{H}_1}(x) - p_{\mathcal{H}_{2,i}}(x)\right| dx$

$\leq \frac{1}{n}\sum_{i=1}^n d_{TV}(\mathcal{N}(0, I_k), \mathcal{H}_{2,i}) \leq \frac{1}{n}\sum_{i=1}^n C\alpha^2 \|L_{B_i} L_{B_i}^T\|_F \leq \frac{1}{n}\sum_{i=1}^n C\alpha^2 \|L_{B_i}\|_F^2$

$\leq \frac{C\alpha^2}{n}\|L\|_F^2 = \frac{C\alpha^2 k}{n}$. The third last inequality follows from Lemma 5. ◀

4.2 Lower bounds for approximating $\|A\|_{1\to p}$ for $1 \leq p \leq 2$

We follow the lower bound template given in Section 4.1.

▶ **Lemma 23.** *For any κ, there exist values s_p such that with probability at least $1 - 1/n$, $\|G_1\|_{1\to p} \leq s_p$ and $\|G_2\|_{1\to p} \geq \kappa s_p$, for $1 \leq p \leq 2$, and $G_1 \sim \mathcal{G}_1$ and $G_2 \sim \mathcal{G}_2[\kappa]$.*

Proof. Recall that from Section 3.1, we know that $\|A\|_{1\to p} = \max_{i\in[n]} \|A_{*,i}\|_p$ which means that it suffices to give bounds on the maximum ℓ_p norm across columns of G_1 and G_2 respectively.

The ℓ_p norm is ζ_p-Lipschitz, where ζ_p is equal to $n^{1/p-1/2}$ in the regime $1 \leq p \leq 2$. For a given vector of standard Gaussians g, the probability that $\|g\|_p$ deviates from $\mathbf{E}[\|g\|_p]$ by more than $\beta\zeta_p\sqrt{\log n}$ is at most $C'e^{-c\beta^2 \log n}$ from Theorem 6 where C' is the constant C from the theorem, which for large enough choice of β can be made smaller than $1/n^2$. By a union bound over all columns, the probability that $\|G_1\|_{1\to p}$ exceeds $\mathbf{E}[\|g\|_p] + \beta\zeta_p\sqrt{\log n}$ is at most $1/n$. On the other hand, consider the perturbed column vector of G_2, which we denote g'. The probability that $\|g'\|_2$ is smaller than $\mathbf{E}[\|g'\|_p] - \beta\sqrt{1+\kappa^2}\zeta_p\sqrt{\log n} = \sqrt{1+\kappa^2}(\mathbf{E}[\|g\|_p] - \beta\zeta_p\sqrt{\log n})$ is at most $1/n^2$ by appropriate choice of β and Theorem 6, from which a lower bound on $\|G_2\|_{1\to p}$ that holds with probability at least $1 - \frac{1}{n^2}$ immediately follows.

Since $\mathbf{E}[\|g\|_p]$ is $\Theta(n^{1/p})$ and the deviations from expectations in upper bounds on $\|G_1\|_{1\to p}$ and lower bounds on $\|G_2\|_{1\to p}$ are asymptotically less than the expectations. ◀

The desired theorem is immediate from Lemma 23, Lemma 22, and Theorem 21 using $\mathcal{D}_1 = \mathcal{G}_{1,n\times n}$, and $\mathcal{D}_2 = \mathcal{G}_2[\kappa]$.

▶ **Theorem 24.** *Suppose $p \in [1, 2]$ and $(\mathcal{S}, f)$ is a $(k, \sqrt{\kappa})$-sketching family for the $1 \to p$ norm where κ is some constant, then $k = \Omega(n)$.*

4.3 Lower bound for approximating $\|A\|_{1\to p}$ for $p > 2$

We follow the lower bound template given in Section 4.1.

Denote $\mathbf{E}[\|g\|_p]$ as η_p. Let $\mathcal{G}_1$ be the distribution over $n \times n$ matrices given by i.i.d. Gaussians, and $\mathcal{G}_2[\alpha, \eta_p]$ be the distribution over $n \times n$ matrices given by taking a Gaussian matrix and adding $\alpha\eta_p$ to a random entry.

Since the proofs are very similar to those in Sections 4.1 and 4.2. We defer them to the full version's Appendix C.1.

▶ **Lemma 25.** *For any κ, there exists s_p such that with probability at least $1 - \frac{1}{n}$, $\|G_1\|_{1\to p} \leq s_p$ and $\|G_2\|_{1\to p} \geq \kappa s_p$, such that $G_1 \sim \mathcal{G}_1$ and $G_2 \sim \mathcal{G}_2[C\kappa, \eta_p]$ for some absolute constant C and $p > 2$.*

▶ **Lemma 26.** *Let L be a linear sketch from $\mathbb{R}^{n\times n} \to \mathbb{R}^k$ and let $\mathcal{D}_i$ be the distribution of $L(x)$ where x is drawn from $\mathcal{G}_i$. Then $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{C'\alpha\eta_p\sqrt{k}}{n}$ for an absolute constant C'.*

The theorem below immediately follows from Lemma 25, Lemma 26 and Theorem 21 using $\mathcal{D}_1 = \mathcal{G}_1$ and $\mathcal{D}_2 = \mathcal{G}_2[C\kappa, \eta_p]$.

▶ **Theorem 27.** *Suppose $(\mathcal{S}, f)$ is a (k, κ)-approximate sketching family for the $1 \to p$ norm for $p > 2$ and some constant κ, then $k = \Omega\left(\frac{n^2}{\eta_p^2}\right)$. In particular, using the fact that η_p is $\Theta(n^{1/p})$ for $p < \infty$ and $\Theta(\sqrt{\log n})$ when $p = \infty$ gives $k = \Omega\left(n^{2-\frac{2}{p}}\right)$ when $p < \infty$ and $k = \Omega\left(\frac{n^2}{\log n}\right)$ when $p = \infty$.*

4.4 Lower bound for approximating $\|A\|_{q\to p}$ when $q \geq 2$ and $p \leq 2$

We use the known lower bound of $\Omega(n^2)$ for sketching the $2 \to 2$ norm from [35] to deduce a lower bound on sketching the $q \to p$ norm for $q \geq 2$ and $p \leq 2$.

▶ **Theorem 28.** *Suppose $q \geq 2$ and $p \leq 2$, and if $(\mathcal{S}, f)$ is a $(k(n), \gamma)$-approximate sketching family for the $q \to p$ norm where γ is some constant, then $k(n) = \Omega(n^2)$.*

Proof. We prove this by showing that if the hypothesis of the theorem statement holds, then the $2 \to 2$ norm can be sketched in $O(k)$ measurements.

Given an $n \times n$ matrix A for which we want to sketch the $2 \to 2$ norm, note that by Lemma 7 there is a $Cn \times n$ matrix L_1 such that $\|L_1A\|_{2\to q^*} = (\frac{1}{\beta}, \beta)\|A\|_{2\to 2}$ for a constant β, and by Lemma 8 $\|L_1A\|_{2\to q^*} = \|A^TL_1^T\|_{q\to 2}$, and another application of Lemma 7 gives us another $Cn \times n$ matrix L_2 for which $\|L_2A^TL_1^T\|_{q\to p} = (\frac{1}{\beta}, \beta)\|A^TL_1^T\|_{q\to 2}$. Note that this means $\|L_2A^TL_1^T\|_{q\to p} = \left(\frac{1}{\beta^2}, \beta^2\right)\|A\|_{2\to 2}$, so we can sketch A by drawing a random L from $\mathcal{D}$ and storing $L(L_2A^TL_1^T)$, which uses $k(Cn)$ measurements and serves as a sketch from which f can be used to estimate $\|A\|_{2\to 2}$ within a constant factor, which means from [35], $k(Cn)$ must be $\Omega(n^2)$, which means $k(n) = \Omega(n^2/C^2) = \Omega(n^2)$. ◀

4.5 Lower bounds for approximating $\|A\|_{q\to p}$ for $p, q \leq 2$ and $p, q \geq 2$

In this section, we show a lower bound on the sketching complexity of $\|A\|_{q\to p}$ where A is a rank r matrix, when both p and q are at most 2. A corresponding lower bound for when p and q are at least 2 follows from Lemma 8. We achieve this by first showing a lower bound on the sketching complexity of $\|A\|_{2\to q}$ and then use Dvoretzky's theorem along with the relation between the $q \to p$ norm and the $p^* \to q^*$ norm to deduce the result.

We show a lower bound for sketching the $2 \to q$ norm using the template from Section 4.1. We use distributions $\mathcal{D}_1 = \mathcal{G}_{1,r\times n}$ and $\mathcal{D}_2[\alpha] = \mathcal{G}_{2,r\times n}\left[\alpha\frac{d}{\sqrt{r}}\right]$, as defined in Section 4.1 where d is $\max\{n^{1/q}, \sqrt{r}\}$.

▶ **Lemma 29.** *There exist values s_q and t_q such that with high probability, $\|G_1\|_{2\to q} \le s_q$ and $\|G_2\|_{2\to q} \ge C\alpha s_q$ for some absolute constant C, for $q > 2$, and $G_1 \sim \mathcal{D}_1$ and $G_2 \sim \mathcal{D}_2[\alpha]$.*

Proof. Let N be a 1/3-net of the Euclidean ball in $\mathbb{R}^r$ with 7^r elements, which exists by Lemma 13. For a fixed $x \in N$, G_1x is distributed as an n-dimensional vector with independent Gaussians, whose q-norm is at most $\beta_1 n^{1/q}$ for some constant β_1 in expectation and exceeds $\beta_1 n^{1/q} + \beta_2\sqrt{r}$ with probability at most $\frac{1}{8^r}$ for appropriate constant β_2, which follows from the q-norm being 1-Lipschitz and Theorem 6. A union bound over all $x \in N$ implies that with probability at least $1 - (7/8)^r$, $\forall x \in N : \|G_1x\|_q \le \beta_1 n^{1/q} + \beta_2\sqrt{r}$.

Then by applying Lemma 12, we conclude that with probability at least $1 - (7/8)^r$, $\|G_1\|_{2\to q} \le \frac{3}{2}(\beta_1 n^{1/q} + \beta_2\sqrt{r}) \le \frac{3}{2}(\beta_1 + \beta_2)d$. On the other hand, the perturbed row of G_2, called g' is distributed as $\sqrt{1 + \alpha^2\frac{d^2}{r}}g$ for a vector of i.i.d. Gaussians g. If we take the unit vector u in the direction of g', then the entry of G_2u corresponding to the perturbed row is concentrated around $\sqrt{1 + \alpha^2\frac{d^2}{r}}\|g\|_2 = \sqrt{r + \alpha^2 d^2}$, which means $\|G_2\|_{2\to q} \ge (1 - o(1))\sqrt{r + \alpha^2 d^2} \ge 0.9\alpha d$ with high probability. ◀

The theorem below immediately follows from Lemma 29, Lemma 22 and Theorem 21.

▶ **Theorem 30.** *Suppose $q \ge 2$ and $(\mathcal{S}, f)$ is a (k, γ)-sketching family for the $2 \to q$ norm of rank r matrices for some constant γ. Then $k = \Omega(nr/d^2)$.*

▶ **Theorem 31.** *Suppose $p, q \le 2$ and $(\mathcal{S}, f)$ is a (k, γ)-sketching family for the $q \to p$ norm of rank r matrices for some constant γ. Then $k = \Omega(nr/d^2)$ where $d = \max\{\sqrt{r}, n^{1/q^*}\}$.*

Proof. For a matrix A, from Lemma 8 we have that $\|A\|_{2\to q^*} = \|A^T\|_{q\to 2}$, and from Lemma 7, we know there is a $Cr \times r$ matrix L_1 such that $\|L_1A^T\|_{q^*\to p} = \Theta(1)\|A\|_{2\to q^*}$. We can use $(\mathcal{S}, f)$ to sketch L_1A^T to obtain an $(O(k), \Theta(1))$-sketching family for the $2 \to q^*$ norm, whose lower bound from Theorem 30 gives us the desired lower bound. ◀

4.6 Lower bounds for approximating $\|A\|_{q\to p}$ for $1 \le q \le 2$ and $p \ge 2$

We prove the desired lower bound using the template from Section 4.1. Let $\mathcal{D}_1$ be a distribution over $n \times n$ matrices where diagonal entries are Gaussians and off-diagonal entries are 0 and let $\mathcal{D}_2[\alpha]$ be a distribution over $n \times n$ matrices where a matrix is drawn from $\mathcal{D}_1$ and $\alpha\sqrt{\log n}$ is added to a random diagonal entry.

▶ **Lemma 32.** *There exists values $s_{p,q}$, $t_{p,q}$ and α such that with probability at least $1 - 1/n$, $\|G_1\|_{q\to p} \le s_{p,q}$ and $\|G_2\|_{q\to p} \ge \kappa s_{p,q}$ for some desired constant factor κ separation, such that $G_1 \sim \mathcal{D}_1$ and $G_2 \sim \mathcal{D}_2[\alpha]$.*

We give the proof of Lemma 32 in the full version's Appendix C.2.

Without loss of generality, we can assume that any sketch of G_1 and G_2 acts on $\mathrm{diag}(G_1)$ and $\mathrm{diag}(G_2)$ respectively. Lemma 26 gives an upper bound of $O(\sqrt{k\log n}/\sqrt{n})$ on the variation distance between k-dimensional sketches of these distributions. Thus, from the variation distance bound, Lemma 32 and Theorem 21, the desired theorem follows.

▶ **Theorem 33.** *Suppose $q \ge 2$ and $(\mathcal{S}, f)$ is a (k, γ)-sketching family for the $q \to p$ norm of rank r matrices for some constant γ, then $k = \Omega(n/\log n)$.*

5 Sketching with large approximation factors

While our results primarily involve constant factor approximations, we give several preliminary results studying large approximation factors for sketching the important cases of the $2 \to q$ norm and $[1, \infty] \to [1, 2]$ norms. Our goal is, given an approximation factor $\alpha(n)$, to give upper and lower bounds on k for a $(k, \alpha(n))$-sketching family for the respective norms. As a shorthand, we will refer to $\alpha(n)$ as α.

5.1 Sketching upper bounds for large approximations of $\|A\|_{2\to q}$

It is sufficient to give a (k, α)-sketching family for the $\infty \to q$ norm. To see why, given an input matrix $A \in \mathbb{R}^{n\times n}$, by Lemma 8 we have that $\|A\|_{2\to q} = \|A^T\|_{q^*\to 2}$. Using Lemma 7, there is a linear map such that this is equal within a constant factor of $\|GA^T\|_{q^*\to 1} = \|AG^T\|_{\infty\to q}$.

▶ **Theorem 34.** *Given a matrix $A \in \mathbb{R}^{n\times n}$, there exists a $(O(\frac{n^2}{\alpha}), \alpha)$-sketching family given by $(\mathcal{S}, f)$ for the $\infty \to q$ norm.*

Proof. Let $B \in \mathbb{Z}^+$ be some positive integer to be chosen later. Let the columns of our sketch matrix S be indexed by sets given by $\{B_i\}_{i=1}^{n/B}$ such that $B_i = ((i-1)B, iB]$. For each column v_{B_i}, we define i.i.d random variables $\{\sigma_{ij}\}_{j=1}^{B}$ such that $\sigma_{ij} = 1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Let the column v_{B_i} be as follows:

$$v_{B_i}[j] = \begin{cases} \sigma_{ij} & \text{for } j \in [(i-1)B, iB] \\ 0 & \text{o/w} \end{cases}$$

We define our linear map $L(A)$ to be $L(A) = AS$. Our function $f : \mathbb{R}^{n/B} \to \mathbb{R}$ simply optimizes over $\{-1, 1\}^{n/B}$ and outputs $\|AS\|_{\infty\to q}$.

Since all $\sigma_{ij} \in \{-1, 1\}$ we have that $f(L(A)) \le \|A\|_{\infty\to q}$ since Sx for $x \in \{-1, 1\}^{n/B}$ has the property that $Sx \in \{-1, 1\}^n$.

We now show a lower bound on $f(L(A))$. To do so, we let T_i denote the column indices of A such that the index is column i in its respective block. We then notice that there exists $i \in [n/B]$ such that $\|A_{*,T_i}\|_{\infty\to q} \ge \frac{B}{n}\|A\|_{\infty\to q}$. We get this by applying the triangle inequality $\|A\|_{\infty\to q} \le \sum_{i=1}^{n/B} \|A_{*,T_i}\|_{\infty\to q}$.

Let i^* be the index that realizes this n/B-approximation to $\|A\|_{\infty\to q}$ and let $\{s_1\}_{i=1}^{n/B}$ be the assignment of signs that realizes the $\infty \to q$ norm of $A_{*,T_{i^*}}$.

$$f(L(A)) \ge \|\sum_{i=1}^{B}\sum_{j=1}^{n/B} s_j A_{*,B_j[i]}\|_q \ge \|\underbrace{\sum_{j=1}^{n/B} s_j A_{*,B_j[i^*]}}_{y} + \underbrace{\sum_{i\neq i^*}^{B}\sum_{j=1}^{n/B} s_j A_{*,B_j[i]}}_{z}\|_q$$

Notice that z is symmetric around the origin and hence we get that $\|y + z + y - z\|_q \le \frac{\|y+z\|_q + \|y-z\|_q}{2}$ which implies that $f(L(A)) \ge \|y + z\|_q \ge \Theta(1)\|y\|_q \ge \frac{n}{B}\|A\|_{\infty\to q}$ with probability at least $\frac{1}{2}$. Thus, we get an $O\left(\frac{n^2}{\alpha}\right)$ space sketch that gives us an α-approximation by setting $B = n/\alpha$. ◀

5.2 Sketching upper bounds for large approximations of $\|A\|_{q\to p}$ for $q \in [1, \infty]$ and $p \in [1, 2]$

We give a description of our sketch followed by the approximation factor. Towards the end of defining our sketch, let $B \in \mathbb{Z}^+$ be some positive integer to be chosen later. Let the rows of our sketch matrix S be indexed by sets given by $\{B_i\}_{i=1}^{n/B}$ such that $B_i = ((i-1)B, iB]$. For

each row v_{B_i}, we define i.i.d random variables $\{\sigma_{ij}\}_{j=1}^{B}$ such that $\sigma_{ij} = 1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Let the row v_{B_i} be as follows:

$$v_{B_i}[j] = \begin{cases} \sigma_{ij} & \text{for } j \in [(i-1)B, iB] \\ 0 & \text{o/w} \end{cases}$$

Our algorithm simply outputs $\|SA\|_{q\to p}$. The proof of the theorem below can be found in the full version's Appendix D.

▶ **Theorem 35.** *Given a matrix $A \in \mathbb{R}^{n\times n}$, there exists an $(\tilde{O}(\frac{n^2}{\alpha^2}), \alpha)$-sketching family given by $(\mathcal{S}, f)$ for the $q \to p$ norm for $p \in [1, 2]$.*

6 Further Directions

One interesting direction is to study the low-rank approximation problem with respect to the $q \to p$ norm. An important open question in the literature is to find input sparsity time low rank approximation algorithms with respect to the $2 \to 2$ norm, and a natural step might be to try this problem with for $q \to p$ norms for certain q and p.

Another interesting problem would be to investigate algorithms for approximate nearest neighbors with respect to the $q \to p$ norm, in light of a question posed by [8] about what metric spaces admit efficient approximate nearest neighbor algorithms, with matrix norms mentioned as an object of interest.

References

1 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.

2 Alexandr Andoni. Nearest neighbor search in high-dimensional spaces. In *the workshop: Barriers in Computational Complexity II*, 2010. URL: `http://www.mit.edu/~andoni/nns-barriers.pdf`.

3 Alexandr Andoni. High frequency moments via max-stability. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 6364–6368, 2017.

4 Alexandr Andoni et al. Eigenvalues of a matrix in the streaming model. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1729–1737. Society for Industrial and Applied Mathematics, 2013.

5 Alexandr Andoni, T. S. Jayram, and Mihai Patrascu. Lower bounds for edit distance and product metrics via poincaré-type inequalities. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 184–192, 2010.

6 Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms via precision sampling. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 363–372. IEEE, 2011.

7 Alexandr Andoni, Robert Krauthgamer, and Ilya P. Razenshteyn. Sketching and embedding are equivalent for norms. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 479–488, 2015.

8 Alexandr Andoni, Huy L Nguyen, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten. Approximate near neighbors for general symmetric norms. In *Proceedings*

of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pages 902–913. ACM, 2017.

9 Alexandr Andoni, Huy L Nguyên, Yury Polyanskiy, and Yihong Wu. Tight lower bound for linear sketches of moments. In *International Colloquium on Automata, Languages, and Programming*, pages 25–32. Springer, 2013.

10 Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 209–218. IEEE, 2002.

11 Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326. ACM, 2012.

12 Aditya Bhaskara and Aravindan Vijayaraghavan. Approximating matrix p-norms. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 497–511. SIAM, 2011.

13 Jaroslaw Blasiok, Vladimir Braverman, Stephen R. Chestnut, Robert Krauthgamer, and Lin F. Yang. Streaming symmetric norms via measure concentration. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 716–729, 2017.

14 Fernando GSL Brandão and Aram W Harrow. Estimating operator norms using covering nets. *arXiv preprint arXiv:1509.05065*, 2015.

15 V. Braverman, S. R. Chestnut, R. Krauthgamer, Y. Li, D. P. Woodruff, and L. F. Yang. Matrix Norms in Data Streams: Faster, Multi-Pass and Row-Order. *ArXiv e-prints*, 2016. `arXiv:1609.05885`.

16 Jop Briët, Fernando Mário de Oliveira Filho, and Frank Vallentin. The positive semidefinite grothendieck problem with rank constraint. In *Automata, Languages and Programming, 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part I*, pages 31–42, 2010.

17 Jop Briët, Oded Regev, and Rishi Saket. Tight hardness of the non-commutative grothendieck problem. *Theory of Computing*, 13(1):1–24, 2017.

18 Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

19 Graham Cormode and S Muthukrishnan. Space efficient mining of multigraph streams. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 271–282. ACM, 2005.

20 Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

21 Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760. ACM, 2015.

22 Aram W Harrow, Ashley Montanaro, and Anthony J Short. Limitations on quantum dimensionality reduction. In *International Colloquium on Automata, Languages, and Programming*, pages 86–97. Springer, 2011.

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

24 Piotr Indyk and David Woodruff. Optimal approximations of the frequency moments of data streams. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 202–208. ACM, 2005.

25 T. S. Jayram. On the information complexity of cascaded norms with small domains. In *2013 IEEE Information Theory Workshop, ITW 2013, Sevilla, Spain, September 9-13, 2013*, pages 1–5, 2013.

26 Thathachar S Jayram and David P Woodruff. The data stream space complexity of cascaded norms. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 765–774. IEEE, 2009.

27 Daniel M Kane, Jelani Nelson, Ely Porat, and David P Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 745–754. ACM, 2011.

28 Daniel M Kane, Jelani Nelson, and David P Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1161–1178. SIAM, 2010.

29 Ashish Khetan and Sewoong Oh. Matrix norm estimation from a few entries. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6427–6436, 2017.

30 Subhash A Khot and Nisheeth K Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative-type metrics into ℓ_1. *Journal of the ACM (JACM)*, 62(1):8, 2015.

31 Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *CoRR*, abs/1602.00061, 2016.

32 Yi Li, Huy L Nguyên, and David P Woodruff. On sketching matrix norms and the top singular vector. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1562–1581. Society for Industrial and Applied Mathematics, 2014.

33 Yi Li and David P. Woodruff. A tight lower bound for high frequency moment estimation with small error. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 623–638, 2013.

34 Yi Li and David P. Woodruff. On approximating functions of the singular values in a stream. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 726–739, 2016.

35 Yi Li and David P Woodruff. Tight bounds for sketching the operator norm, schatten norms, and subspace embeddings. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 60. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

36 Yi Li and David P. Woodruff. Embeddings of schatten norms with applications to data streams. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, pages 60:1–60:14, 2017.

37 Jirı Matoušek. Lecture notes on metric embeddings. Technical report, Technical report, ETH Zürich, 2013.

38 Cameron Musco, Praneeth Netrapalli, Aaron Sidford, Shashanka Ubaru, and David P. Woodruff. Spectrum approximation beyond fast matrix multiplication: Algorithms and hardness. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 8:1–8:21, 2018.

39 Assaf Naor, Oded Regev, and Thomas Vidick. Efficient rounding for the noncommutative grothendieck inequality. *Theory of Computing*, 10:257–295, 2014.

40 Eric Price and David P. Woodruff. Applications of the shannon-hartley theorem to data streams and sparse recovery. In *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, pages 2446–2450, 2012.

41 Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

42 Zhao Song, David P. Woodruff, and Peilin Zhong. Low rank approximation with entrywise l_1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 688–701, 2017.

43 Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

44 Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of tr(f(a)) via stochastic lanczos quadrature, 2016. URL: `http://www-users.cs.umn.edu/~saad/PDF/ys-2016-04.pdf`.

45 Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

46 Andreas J. Winter. Quantum and classical message identification via quantum channels. *Quantum Information & Computation*, 5(7):605–606, 2005.

47 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

A Proofs from Section 2

Proof of Lemma 9. For any x that is unit according to ℓ_1,

$$\begin{aligned}\|Ax\|_p &= \|A_{*,1}x_1 + A_{*,2}x_2 + \ldots + A_{*,n}x_n\|_p \\ &\leq \|A_{*,1}\|_p|x_1| + \|A_{*,2}\|_p|x_2| + \ldots + \|A_{*,n}\|_p|x_n| \leq \max_{i\in[n]}\{\|A_{*,i}\|_p\}\end{aligned}$$

where the last inequality is because $|x_i|$ give a convex combination and is achieved for $x = e_{i^*}$ where $i^* = \arg\max_i\{\|A_{*,i}\|_p\}$. ◀

Proof of Lemma 10. For any x such that there is a coordinate x_j that is strictly between 1 or -1, let ε be $\min\{1 - x_j, x_j + 1\}$, consider

$$\begin{aligned}\|Ax\|_p &= \|A_{*,j}x_j + \sum_{i\neq j} A_{*,i}x_i\|_p \\ &\leq \left(\frac{1+x_j}{2}\right)\|A_{*,j} + \sum_{i\neq j} A_{*,i}x_i\|_p + \left(\frac{1-x_j}{2}\right)\|-A_{*,j} + \sum_{i\neq j} A_{*,i}x_i\|_p\end{aligned}$$

where the inequality is due to the triangle inequality. Since $\|Ax\|_p$ is at most a convex combination of the p-norms after replacing x_j with 1 or -1, we can make x_j one of 1 or -1 without decreasing the p-norm. ◀

Proof of Lemma 12. Pick x^* on the unit ball such that $\|Ax^*\|_{\mathcal{Y}} = \|A\|_{\mathcal{X}\to\mathcal{Y}}$. There is $x \in N$ such that $\|x^* - x\|_{\mathcal{X}} < \varepsilon$, which means

$$\|A(x^* - x)\|_{\mathcal{Y}} \leq \|A\|_{\mathcal{X}\to\mathcal{Y}}\|x - x^*\|_{\mathcal{X}} < \varepsilon\|A\|_{\mathcal{X}\to\mathcal{Y}}$$

On the other hand,

$$\|A(x^* - x)\|_{\mathcal{Y}} \geq \|Ax^*\|_{\mathcal{Y}} - \|Ax\|_{\mathcal{Y}} \geq \|A\|_{\mathcal{X}\to\mathcal{Y}} - \|Ax\|_{\mathcal{Y}}$$

and hence

$$\begin{aligned}\|A\|_{\mathcal{X}\to\mathcal{Y}} - \|Ax\|_{\mathcal{Y}} &< \varepsilon\|A\|_{\mathcal{X}\to\mathcal{Y}} \\ \|A\|_{\mathcal{X}\to\mathcal{Y}} < \frac{\|Ax\|_{\mathcal{Y}}}{1-\varepsilon} &\leq \frac{1}{1-\varepsilon}\max_{x\in N}\|Ax\|_{\mathcal{Y}}\end{aligned}$$ ◀

Proof of Lemma 13. For x in a normed space $\mathcal{X}$, we use the notation $B_x(r)$ to denote $\{y : \|x-y\|_{\mathcal{X}} < r\}$, the ball of radius r around x.

Start with an empty set N and while there is a point x in the unit ball B that has distance at least ε to every element in N, pick x and add it to N. This process terminates when every $x \in B$ has distance less than ε to some element in N, thereby terminating with N as an ε-net. We claim that the size of N meets the desired bound.

By construction, any y and y' in N are at least ε apart, which means $\mathcal{B} = \{B_x(\varepsilon/2) : x \in N\}$ is a collection of disjoint sets and note that

$$\bigcup_{S\in\mathcal{B}} S \subseteq B_0(1+\varepsilon/2)$$

By disjointness

$$\mathrm{Vol}\left(\bigcup_{S\in\mathcal{B}} S\right) = \sum_{S\in\mathcal{B}} \mathrm{Vol}(S) = |N|\mathrm{Vol}(B_0(\varepsilon/2))$$

where $\mathrm{Vol}(S)$ is the volume of S according to the Lebesgue measure.

And thus, we obtain

$$\begin{aligned}
|N| &= \frac{\mathrm{Vol}\left(\bigcup_{S\in\mathcal{B}} S\right)}{\mathrm{Vol}(B_0(\varepsilon/2))} \\
&\leq \frac{\mathrm{Vol}(B_0(1+\varepsilon/2))}{\mathrm{Vol}(B_0(\varepsilon/2))} \\
&= \left(\frac{1+\varepsilon/2}{\varepsilon/2}\right)^n \\
&= \left(\frac{2+\varepsilon}{\varepsilon}\right)^n
\end{aligned}$$

which concludes the proof. ◀

B Missing proofs from Section 3

Proof of Lemma 16. Draw $c\log n$ matrices $S_1, S_2, \ldots, S_{c\log n}$ from $\mathcal{D}$ independently where c is a constant to be determined later. We define

$$S := \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_{c\log n} \end{bmatrix}$$

$$g(Sx) := \mathsf{median}\{f(S_1x), f(S_2x), \ldots, f(S_{c\log n}x)\}$$

Let's analyze the probability that $g(Sx)$ falls outside $L_x = \left(\frac{1}{2}\|x\|_p, 2\|x\|_p\right)$. In order for that to happen, more than half of $f(S_1x), \ldots, f(S_{c\log n}x)$ must lie outside L_x, and this happens to each $f(S_ix)$ with probability at most $\frac{1}{3}$. Using Hoeffding's inequality, we know

$$\mathbf{Pr}[g(Sx) \notin L] \leq 2\exp\left(-\frac{c\log n}{72}\right)$$

which for appropriate choice of c can be bounded by $\frac{1}{n^2}$.

For a matrix A with n columns, a union bound tells us that for all i, $g(SA_{*,i})$ falls in $L_{A_{*,i}}$ with probability at least $1-\frac{1}{n}$. Combined with Lemma 3.1, it follows that $h(SA) := \max_i g(SA_{*,i})$ is a 2-approximation to $\|A\|_{1\to p}$ with probability at least $1-\frac{1}{n}$. ◀

C Missing Proofs from Section 4

C.1 Missing Proofs from Section 4.3

Proof of Lemma 25. We denote $C\kappa$ as α and set the exact value of α in the end of the proof. For a fixed pair i, j let us denote the perturbation term $\alpha\eta_p e_i e_j^\top$ as E_{ij}. Recall that from section 3.1, we know that $\|A\|_{1\to p} = \max_{i\in[n]} \|A_{*,i}\|_p$ which means that it suffices to give bounds on the maximum ℓ_p norm across columns of G_1 and G_2 respectively.

Since the ℓ_p norm is 1-Lipschitz for any $p \geq 2$, we can apply Theorem 6 to show concentration around the expectation for $\|G_{*,i}\|_p$ for any column i of a matrix G of i.i.d Gaussian entries. Hence we have that for any column i, and some positive constant λ

$$\mathbf{Pr}\left[\|G_{*,i}\|_p \geq \lambda \mathbf{E}[\|G_{*,i}\|_p]\right] \leq C\exp(-c\lambda^2 \mathbf{E}[\|G_{*,i}\|_p]^2)$$

Letting g be an n-dimensional vector of i.i.d Gaussians, since we know $\mathbf{E}[\|g\|_p] = \Omega(\sqrt{\log n})$, there exists appropriate constant β such that for any column i of G_1 we have that $\|(G_1)_{*,i}\|_p$ is less than $\beta\mathbf{E}[\|g\|_p]$ with probability at least $1 - \frac{1}{n^2}$. By a union bound over all columns, the probability that $\|G_1\|_{1\to p} \leq \beta\mathbf{E}[\|g\|_p]$ is at least $1 - \frac{1}{n}$.

For a matrix $G_2 = G + E_{ij}$ drawn from $\mathcal{G}_2[\alpha, \eta_p]$, we know that the perturbed column j has norm at least $\alpha\eta_p - \|G_{*,i}\|_p$, which satisfies $(\alpha - \beta)\mathbf{E}[\|g\|_p] \leq \|G_2\|_{1\to p}$. Setting $\alpha \geq (\kappa + 1)\beta$ gives us the desired result. ◀

Proof of Lemma 26. Recall perturbation term $\alpha\eta_p e_i e_j^\top$ was referred to as E_{ij}. Just as in Lemma 22, we can think of L as a $k \times n^2$ matrix that acts on a sample from $\mathcal{G}_1$ or $\mathcal{G}_2[\alpha]$ as though it were an n^2-dimensional vector. Without loss of generality, we can assume that the rows of L are orthonormal, since as before we can always perform a change of basis in post-processing. Thus, the distribution $\mathcal{D}_1$ is the same as $\mathcal{N}(0, I_k)$. For fixed i, j, the distribution of $L(G + E_{ij})$ is Gaussian with mean vector $L(E_{ij})$ (the ij^{th} column of the $k \times n^2$ matrix L scaled by $\alpha\eta_p$) and covariance I_k because of the following.

$$\begin{aligned}
\mathsf{Cov}(L(G + E_{ij})) &= \mathbf{E}\left[\left(L(G + E_{ij}) - \mathbf{E}\left[L(G + E_{ij})\right]\right)^\top \left(L(G + E_{ij}) - \mathbf{E}\left[L(G + E_{ij})\right]\right)\right] \\
&= \mathbf{E}\left[\left(L(G) - \mathbf{E}\left[L(G)\right]\right)^\top \left(L(G) - \mathbf{E}\left[L(G)\right]\right)\right] \\
&= \mathsf{Cov}_{G\sim\mathcal{N}(0,I_n)}(G) = I_k
\end{aligned}$$

Thus, $\mathcal{D}_2$ is the distribution of picking a random i, j and drawing a matrix from $\mathcal{N}(L(E_{ij}), I_k)$.

We now analyze the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ and get the desired bound from a chain of inequalities.

$$\begin{aligned}
d_{TV}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{x\in\mathbb{R}^k} |p_{\mathcal{D}_1}(x) - p_{\mathcal{D}_2}(x)|dx \\
&= \frac{1}{2}\int_{x\in\mathbb{R}^k} \left|\sum_{i,j} \frac{1}{n^2} p_{\mathcal{D}_1}(x) - \frac{1}{n^2} p_{\mathcal{N}(L(E_{ij}),I_k)}(x)\right| dx \\
&\leq \frac{1}{n^2}\sum_{i,j} \frac{1}{2}\int_{x\in\mathbb{R}^k} \left|p_{\mathcal{D}_1}(x) - p_{\mathcal{N}(L(E_{ij}),I_k)}\right| dx \\
&= \frac{1}{n^2}\sum_{i,j} d_{TV}(\mathcal{D}_1, \mathcal{N}(L(E_{ij}), I_k))
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i,j} d_{TV}(\mathcal{N}(0,I_k),\mathcal{N}(L(E_{ij}),I_k)) \\
&\le \frac{1}{n^2}\sum_{i,j} C'\alpha\eta_p\|L_{*,ij}\|_2 && \text{[from lemma 5]} \\
&= \frac{C'\alpha\eta_p}{n^2}\|L\|_{1,2} \\
&\le \frac{C'\alpha\eta_p}{n^2}\cdot n\|L\|_F = C'\alpha\eta_p\cdot\frac{\sqrt{k}}{n} && \text{[by Cauchy-Schwarz]}
\end{aligned}
$$

◀

C.2 Missing Proofs from Section 4.6

Proof of Lemma 32. We claim that for a diagonal matrix D, $\arg\max_{\|x\|_q=1}\|Dx\|_p$ is achieved when x is one of the e_i standard basis vectors e_i. To see this,

$$\|Dx\|_p^p = \sum_{i=1}^n |d_{ii}x_i|^p = \sum_{i=1}^n |d_{ii}|^p(|x_i|^q)^{p/q} \le \sum_{i=1}^n |d_{ii}|^p|x_i|^q \le \max_i |d_{ii}|^p$$

which is achieved by picking $x = e_{i^*}$ where choice of $i = i^*$ maximizes d_{ii}.

Thus, to analyze the $q \to p$ norm of G_1, it suffices to analyze $\max_{x\in\{e_i\}}\|G_1x\|_p$, which is the same as $\|g\|_\infty$ where g is a vector of i.i.d. Gaussians. We can extract from the proof of Lemma 25 that $\|g\|_\infty$ is upper bounded by $\beta\sqrt{\log n}$ with probability at least $1-\frac{1}{n^2}$.

On the other hand, if the perturbation is at index (i,i) and we pick $\alpha = \kappa(\beta+1)$, then $\|G_2e_i\|_p$ is at least $\kappa\beta\sqrt{\log n}$ with probability at least $1-\frac{1}{n^2}$ implying the desired separation. ◀

D General approximation factors α

D.1 Sketching Matrix Construction and Upper Bounds

Let us first define our sketch and then analyze its performance. For the sketch S, we group the rows of A into $\frac{n}{\alpha^2}$ groups of size α^2. We label the groups by $B_1,\ldots,B_{n/\alpha^2}$ and let $\sigma_{1i},\ldots,\sigma_{\alpha^2 i}$ be ±1 i.i.d random variables with equal probability for block B_i. Notice then that the i^{th} row of SA given by $(SA)_{i,*}$ is:

$$(SA)_{i,*} \triangleq \sum_{j\in B_i}\sigma_{ji}A_{i,*}$$

To analyze the performance of this sketch, we will need a helpful inequality describing the behavior of a random signed sums of reals.

▶ **Theorem 36** (Khintchine's Inequality, [20]). *Let $\{x_i\}_{i=1}^n \in \mathbb{R}$ be reals and let $\{\mathbf{s_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ be i.i.d ±1 random variables with equal probability and let $0 < t < \infty$, we then have:*

$$A_p\sqrt{\sum_{i=1}^n x_i^2} \le \boldsymbol{E}\left[\left|\sum_{i=1}^n \mathbf{s_i}\mathbf{x_i}\right|^p\right]^{1/p} \le B_p\sqrt{\sum_{i=1}^n x_i^2}$$

For some constants A_p, B_p that only depend on p.

Also recall that by Jensen's inequality, we can relate two norms of a vector $x \in \mathbb{R}^n$.

▶ Remark. For two positive reals, $p \geq q > 1$ and for a vector $x \in \mathbb{R}^n$ we have that: $\|x\|_p \leq n^{\frac{1}{q}-\frac{1}{p}} \|x\|_q$

We then have the following theorems describing the sketching complexity of the sketch S for $1 \leq p \leq 2$ and for $p > 2$.

▶ **Theorem 37.** *For any $1 \leq p \leq 2$ and for the maximizer $x \in \mathbb{R}^n$ of $\|A\|_{q\to p}$ the sketch S defined earlier where each block B_i has size B has the property that*

$$\Theta(1)\frac{1}{B^{1-\frac{1}{p}}} \|SAx\|_p \leq \|Ax\|_p \leq \Theta(1)B^{\frac{1}{p}-\frac{1}{2}} \|SAx\|_p$$

with probability at least $\frac{99}{100}$

Proof. Let us first show the first inequality in the theorem statement.

For some coordinate $1 \leq i \leq \frac{n}{B}$:

$$|(SAx)_i|^p = \left|\sum_{j\in B_i} \sigma_j (Ax)_j\right|^p \leq \left(\sum_{j\in B_i} |(Ax)_j|\right)^p$$

By Remark D.1 relating $\|\cdot\|_1$ and $\|\cdot\|_p$

$$\leq B^{p-1} \sum_{j\in B_i} |(Ax)_j|^p$$

$$\therefore \|(SAx)_i\|_p = \left(\sum_{i=1}^{n/B} |(SAx)_i|^p\right)^{1/p} \leq B^{1-\frac{1}{p}} \|Ax\|_p$$

Notice that the first inequality holds irrespective of the vector x, it holds for all vectors. Now let us show the second inequality of the theorem statement.

For some coordinate $1 \leq i \leq \frac{n}{B}$:

$$\left(\sum_{j\in B_i} (Ax)_j^p\right)^{1/p} \leq B^{\frac{1}{p}-\frac{1}{2}} \left(\sum_{j\in B_i} (Ax)_j^2\right)^{1/2} \quad \text{[By Remark D.1]} \quad [1]$$

$$\leq \Theta(1)B^{\frac{1}{p}-\frac{1}{2}} \mathbf{E}\left[\left|\sum_{j\in B_i} \sigma_j (Ax)_j\right|^p\right]^{1/p} \quad \text{[By Khintchine's Ineq.]} \quad [2]$$

$$\therefore \sum_{i=1}^{n/B} \sum_{j\in B_i} (Ax)_j^p = \|Ax\|_p^p \leq \Theta(1)B^{p\left(\frac{1}{p}-\frac{1}{2}\right)} \mathbf{E}\left[\|SAx\|_p^p\right]$$

Notice that the second inequality of the theorem statement follows by Markov's inequality.

Notice that the success probability of line [2] is constant for each block. To get constant success probability over the entire set of blocks, we construct $O(\log(n))$ i.i.d copies of each block B_i given by $\{B_i^j\}_{i=1}^{O(\log(n))}$. We then pick j such that it is the index realizing the quantity $\mathsf{median}_{j\in[O(\log(n))]} \|(S_j Ax)_i\|_p$ where S_j corresponds the sketch with the j^{th} copy of the blocks. Then, by standard concentration bounds, we can get $1 - \frac{1}{n/B}$ success probability for each set of blocks B_i and then union bound over the $\frac{n}{B}$ blocks giving us constant success probability. ◀

▶ **Theorem 38.** *For any $p > 2$ and for the maximizer $x \in \mathbb{R}^n$ of $\|A\|_{q\to p}$ the sketch S defined earlier where each block B_i has size B has the property that*

$$\Theta(1)\frac{1}{B^{1-\frac{1}{p}}}\|SAx\|_p \leq \|Ax\|_p \leq \Theta(1)\|SAx\|_p$$

The proof for Theorem 38 is the same as that for Theorem 37 except that there is no dilation while upper bounding the $\|Ax\|_p$ with the 2-norm in line [1] of the proof.

Notice that the above theorems imply that the sketch S is a $\sqrt{B}$-approximation when $0 \leq p \leq 2$ and a $B^{1-\frac{1}{p}}$-approximation when $p > 2$ because it states that the sketch is stretching $\|Ax\|_p^p$ by at most some factor and dilating it by at most some factor and hence the approximation ratio is simply the product of these factors.