# The Algorithmics Column

by


Gerhard J Woeginger

Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
gwoegi@win.tue.nl

# CHAINING INTRODUCTION WITH SOME COMPUTER SCIENCE APPLICATIONS

Jelani Nelson*

## Contents



*Harvard University. minilek@seas.harvard.edu. Supported by NSF CAREER award CCF-1350670, NSF grant IIS-1447471, ONR Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

# 1 What is chaining?

Consider the problem of bounding the maximum of a collection of random variables. That is, we have some collection $(X_t)_{t\in T}$ and want to bound $\mathbb{E}\sup_{t\in T} X_t$, or perhaps we want to say this sup is small with high probability (which can be achieved by bounding $\mathbb{E}\sup_{t\in T} |X_t|^p$ for large $p$ and applying Markov's inequality).

Such problems show up all the time in probabilistic analyses, including in computer science, and the most common approach is to combine tail bounds with union bounds. For example, to show that the maximum load when throwing $n$ balls into $n$ bins is $O(\log n/\log\log n)$, one defines $X_t$ as the load in bin $t$, proves $\mathbb{P}(X_t > C\log n/\log\log n) \ll 1/n$, then performs a union bound to bound $\sup_t X_t$. Or when analyzing the update time of a randomized data structure on some sequence of operations, one argues that no operation takes too much time by understanding the tail behavior of $X_t$ being the time to perform operation $t$, then again performs a union bound to control $\sup_t X_t$.

Most succinctly, chaining methods leverage statistical dependencies between a (possibly infinite) collection of random variables *to beat this naive union bound*.

The origins of chaining began with Kolmogorov's continuity theorem from the 1930s (see Section 2.2, Theorem 2.8 of [21]). The point of this theorem was to understand conditions under which a stochastic process is continuous. That is, consider a random function $f : \mathbb{R} \to X$ where $(X, d)$ is a metric space. Assume the distribution over $f$ satisfies the property that for some $\alpha, \beta > 0$, $\mathbb{E}|f(x) - f(y)|^\alpha = O(|x - y|^{1+\beta})$ for all $x, y \in \mathbb{R}$. Kolmogorov proved that for any such distribution, one can couple with another distribution over functions $\tilde{f}$ such that $\forall x \in \mathbb{R}$, $\mathbb{P}(f(x) = \tilde{f}(x)) = 1$, and furthermore $\tilde{f}$ is continuous. For the reader interested in seeing proof details, see for example [29, Section A.2].

Since Kolmogorov's work, the scope of applications of the chaining methodology has widened tremendously, due to contributions of many mathematicians, including Dudley, Fernique, and very notably Talagrand. See Talagrand's treatise [29] for a description of many impressive applications of chaining in mathematics. See also Talagrand's STOC 2010 paper [28]. Note that [29] is not exhaustive, and additional applications are posted on the arXiv on a regular basis.

# 2 Applications in computer science

Several applications are given in [30, Section 1.2.2]. I will repeat some of those here, as well as some other ones.

## 2.1 Random matrices and compressed sensing

Consider a random matrix $M \in \mathbb{R}^{m \times n}$ from some distribution. A common task is to understand the behavior of the largest singular value of $M$. Note $\|M\| = \sup_{\|x\|_2 = \|y\|_2 = 1} x^T M y$, so the goal is to understand the supremum of the random variables $X_t = t_1^T M t_2$ for $t \in T = B_{\ell_2^m} \times B_{\ell_2^n}$. Indeed, for many distributions one can obtain asymptotically sharp results via chaining.

Understanding singular values of random matrices has been important in several areas of computer science. Close to my own heart are in compressed sensing and randomized linear algebra algorithms. For the latter, a relevant object is a *subspace embedding*; these are objects used in algorithms for fast regression, low-rank approximation, and a dozen other applications (see [31]). Analyses then boil down to understanding the largest singular value of $M = (\Pi U)^T(\Pi U) - I$. In compressed sensing, where the goal is to approximately recover a nearly sparse signal $x$ from few linear measurements $Sx$ (the measurements are put as rows of the matrix $S$), analyses again boil down to bounding the operator norm of the same $M$, but for all $U$ simultaneously that can be formed from choosing $k$ columns from some basis that $x$ is sparse in.

## 2.2 Empirical risk minimization

This example is taken from [30]. In machine learning one often is given some data, drawn from some unknown distribution, and a loss function $\mathcal{L}$. Given some family of distributions parameterized by some $\theta \in \Theta$, the goal is to find some $\theta^*$ which explains the data the best, i.e.

$$\theta^* = \underset{\theta \in \Theta}{argmin}\, \mathbb{E}\, \mathcal{L}(\theta, X). \tag{1}$$

The expectation is taken over the distribution of $X$. We do not know $X$, however, and only have i.i.d. samples $X_1, \ldots, X_n$. Thus a common proxy is to calculate

$$\hat{\theta} = \underset{\theta \in \Theta}{argmin}\, \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}(\theta, X_k).$$

We would like to argue that $\hat{\theta}$ is a nearly optimal minimizer for the actual problem (1). For this to be true, it is sufficient that $\sup_\theta X_\theta$ is small, where one ranges over all $\theta \in \Theta$ with

$$X_\theta = \left| \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}(\theta, X_k) - \mathbb{E}\, \mathcal{L}(\theta, X) \right|.$$

## 2.3 Dimensionality reduction

In Euclidean dimensionality reduction, such as in the Johnson-Lindenstrauss lemma, one is given a set of vectors $P \subset \ell_2^n$, and wants that a (usually random) matrix $\Pi$ satisfies

$$\forall y, z \in P,\ (1-\varepsilon)\|y-z\|_2^2 \le \|\Pi y - \Pi z\|_2^2 \le (1+\varepsilon)\|y-z\|_2^2. \tag{2}$$

This is satisfied as long as $\sup_{y,z} X_{y,z} \le \varepsilon$, where

$$X_{y,z} = |\frac{1}{\|y-z\|_2^2}\|\Pi y - \Pi z\|_2^2 - 1|,$$

where $y, z$ ranges over all pairs of distinct vectors in $P$. Gordon's theorem [15] states that a $\Pi$ with i.i.d. gaussian entries ensures this with good probability as long as it has $\gtrsim (g^2(T)+1)/\varepsilon^2$ rows, where $g(T)$ is the *gaussian mean width* of $T$ and $T$ is the set of normalized differences of vectors in $P$. Later works gave sharper analysis, and also extended to other types of $\Pi$, all using chaining [19, 24, 2, 6, 9, 25].

Another application of chaining in the context dimensionality reduction was in regard to nearest neighbor (NN) preserving embeddings [17]. In this problem, one is given a database $X \subset \ell_2^d$ of $n$ points and must create a data structure such that for any query point $q \in \mathbb{R}^d$, one can quickly find a point $x \in X$ such that $\|q - x\|_2$ is nearly minimized. Of course, if *all* distances are preserved between $q$ and points in $X$, this suffices to accomplish our goal, but it is more powerful than what is needed. It is only needed that the distance from $q$ to its nearest neighbor does not increase too much, and that the distances from $q$ to much farther points do not shrink too much (to fool us into thinking that they are approximate nearest neighbors). An embedding satisfying such criteria is known as a *NN-preserving embedding*, and [17] used chaining methods to show that certain "nice" sets $X$ have such embeddings into low dimension. Specifically, the target dimension can be $O(\Delta^2\varepsilon^{-2}\frac{\gamma_2(X)}{diam(X)})^2$, where $\Delta$ is the aspect ratio of the data and $\gamma_2$ is a functional defined by Talagrand (more on that later). All we will say now is that $\gamma_2(X)$ is always $O(\sqrt{\log \lambda_X})$, where $\lambda_X$ is the doubling constant of $X$ (the maximum number of balls of radius $r/2$ required to cover any radius-$r$ ball, over all $r$).

## 2.4 Data structures and streaming algorithms

The potential example to data structures was already mentioned in the previous section. To make it more concrete, consider the following streaming data structural problem in which one sees a sequence $p_1, \ldots, p_m$ with each $p_k \in \{1, \ldots, n\}$. For example, when monitoring a search query stream, $p_k$ may be a word in a dictionary of size $n$. The goal of the *heavy hitters* problem is to identify words that occur frequently in the stream. Specifically, if we let $f_i$ be the number of occurrences of $i \in [n]$ in the stream, in the $\ell_2$ heavy hitters problem the goal is to find all $i$ such that $f_i^2 \geq \varepsilon \sum_i f_i^2$ (think of $\varepsilon$ as some given constant). The CountSketch of Charikar, Chen, and Farach-Colton solves this problem using $O(\log n)$ machine words of memory.

A recent work of [5] provides a new algorithm that solves the same problem using only $O(\log \log n)$ words of memory, and even more recently it has been shown how to achieve the optimal $O(1)$ words of memory [4]. These are randomized algorithms that maintain certain random variables in memory that evolve over time, and their analyses require controlling the largest of their deviations. Without getting into technical details here, we describe a related streaming problem: $\ell_2$ estimation. The goal here is to use small memory while, after any query, being able to output an estimate $Q$ satisfying $\mathbb{P}(|Q - \|f\|_2| > \varepsilon\|f\|_2) < 1/3$ (the probability is over the randomness used by the algorithm). It turns out this problem can be solved in $O(1/\varepsilon^2)$ words of memory by a randomized data structure known as the "AMS sketch" [3]. The failure probability can be decreased to $\delta$ by running $\Theta(\log(1/\delta))$ instantiations of the algorithm in parallel with independent randomness, then returning the median estimate of $\|f\|_2$ during a query. This yields space $O(\varepsilon^{-2}\log(1/\delta))$ words, which is optimal [18].

Recently the following question has been studied: what if we want to track $\|f\|_2$ *at all times*? Recalling the stream contains $m$ updates, one could do as above and set $\delta < 3/m$ and union bound, so with an $O(\varepsilon^{-2}\log m)$-space algorithm, with probability $2/3$ all queries throughout the entire stream are correct. The work [16] showed this bound can be asymptotically improved when the number of distinct indices in the stream and $1/\varepsilon$ are both subpolynomial in $m$. This restriction was removed in subsequent works [5, 4].

## 2.5 Random walks on graphs

Ding, Lee, and Peres [11] a few years ago gave the first *deterministic* constant-factor approximation algorithm to the cover time of a random graph. Their work showed that the cover time of any connected graph is, up to a constant, equal to the supremum of a certain collection of random variables depending on that graph: the *gaussian free field*. This is a collection of gaussian random variables whose covariance structure is given by the effective resistances between the graph's vertices. Work of Talagrand (the "majorizing measures theory") and Fernique have provided us with tight, up to a constant factor, upper and lower bounds for the expected supremum of a collection of random variables. Furthermore, these bounds are constructive and efficient. See also the works [23, 8, 32] for more on this topic.

## 2.6 Dictionary learning

In *dictionary learning* one assumes that some data of $p$ samples, the columns of some matrix $Y \in \mathbb{R}^{n \times p}$, is (approximately) sparse in some unknown "dictionary". That is, $Y = AX + E$ where $A$ is unknown, $X$ is sparse in each column, and $E$ is an error matrix. If $E = 0$, $A$ is square, and $X$ has i.i.d. entries with $s$ expected non-zeroes per column, with the non-zeroes being subgaussian, then Spielman, Wang, and Wright gave the first polynomial-time algorithm which provably recovers $A$ (up to permutation and scaling of its columns) using polynomially many samples. Their proof required $O(n^2 \log^2 n)$ samples, but they conjectured $O(n \log n)$ should suffice.

It was recently shown that their precise algorithm needs roughly $n^2$ samples, but $O(n \log n)$ does suffice for a slight variant of their algorithm. As per [27], the analysis of the latter result boiled down to bounding the supremum of a collection of random variables. See [22, 1, 7].

## 2.7 Error-correcting codes

A $q$-ary linear error-correcting code $C$ is such that the codewords are all vectors of the form $xM$ for some row vector $x \in \mathbb{F}_q^m$ and $M \in \mathbb{F}_q^{m \times n}$. $M$ is called the "generator matrix". Such a code is *list-decodable* up to some radius $R$, if, informally, if one arbitrarily corrupts any codeword $C$ in at most an $R$-fraction of coordinates to obtain some $C'$, then the *list* of candidate codewords in $C$ which could have arisen in this way (i.e. are within radius $R$ of $C'$) is small.

Recent work of Rudra and Wootters [26] showed, to quote them, that "any $q$-ary code with sufficiently good distance can be randomly punctured to obtain, with high probability, a code that is list decodable up to radius $1 - 1/q - \varepsilon$ with near-optimal rate and list sizes". A "random puncturing" means simply to randomly sample some number of columns of $M$ to form a random matrix $M'$, which is the generator matrix for the new "punctured" code. Their proof relies on chaining.

In the remainder, we show the details of how chaining works, we play with a toy example (bounding the gaussian mean width of the $\ell_1$ ball in $\mathbb{R}^n$), then describe an application of chaining to a real computer science problem: Euclidean dimensionality reduction.

# 3 A case study: (sub)gaussian processes

To give an introduction to chaining, I will focus our attention on a concrete scenario. Suppose we have a bounded (but possibly infinite) collection of vectors

$T \subset \mathbb{R}^n$. Furthermore, let $g \in \mathbb{R}^n$ be a random vector with its entries being independent, mean zero, and unit variance gaussians. We will consider the collection of variables $(X_t)_{t \in T}$ with $X_t$ defined as $\langle g, t \rangle$. In what follows, we will only ever use one property of these $X_t$:

$$\forall s, t \in T,\ \mathbb{P}(|X_s - X_t| > \lambda) \lesssim e^{-\lambda^2/(2\|s-t\|_2^2)}. \tag{3}$$

This provides us with some understanding of the dependency structure of the $X_t$. In particular, if $s, t$ are close in $\ell_2$, then it's very likely that the random variables $X_s$ and $X_t$ are also close.

Why does this property hold? Well,

$$X_s - X_t = \langle g, s - t \rangle = \sum_{i=1}^{n} g_i \cdot (s - t)_i.$$

We then use the property that adding independent gaussians yields a gaussian in which the variances add. If you haven't seen that fact before, it follows easily from looking at the Fourier transform of the gaussian pdf. Adding independent random variables convolves their pdfs, which pointwise multiplies their Fourier transforms. Since the Fourier transform of a gaussian pdf is a gaussian whose variance is inverted, it then follows that summing independent gaussians gives a gaussian with summed variances. Thus $X_s - X_t$ is a gaussian with variance $\|s-t\|_2^2$, and (3) then follows by tail behavior of gaussians. Note (3) would hold for subgaussian distributions too, such as for example $g$ being a vector of independent uniform $\pm 1$ random variables.

Now I will present four approaches to bounding $g(T) := \mathbb{E}_g \sup_{t \in T} X_t$. These approaches will be gradually sharper. For simplicity I will assume $|T| < \infty$, although it is easy to circumvent this assumption for methods 2, 3, and 4.

## 3.1 Method 1: union bound

Remember that, in general for a scalar random variable $Z$,

$$\mathbb{E}\,|Z| = \int_0^\infty \mathbb{P}(Z > u)du.$$

Let $\rho_X(T)$ denote the diameter of $T$ under norm $X$. Then

$$
\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &= \int_0^\infty \mathbb{P}(\sup_{t\in T} X_t > u)du \\
&\le \int_0^{2\rho_{\ell_2}(T)\sqrt{2\log|T|}} \overbrace{\mathbb{P}(\sup_{t\in T} X_t > u)}^{\le 1} du + \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}} \mathbb{P}(\sup_{t\in T} X_t > u)du \\
&\le \rho_{\ell_2}(T)\sqrt{2\log|T|} + \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}}^\infty \sum_{t\in T}\mathbb{P}(X_t > u)du \text{ (union bound)} \\
&\le \rho_{\ell_2}(T)\sqrt{2\log|T|} + |T|\cdot \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}}^\infty e^{-u^2/(2\rho_{\ell_2}(T)^2)}du \\
&= \rho_{\ell_2}(T)\sqrt{2\log|T|} + \rho_{\ell_2}(T)\cdot|T|\cdot\int_{\sqrt{2\log|T|}}^\infty e^{-v^2/2}dv \text{ (change of variables)} \\
&\lesssim \rho_{\ell_2}(T)\cdot\sqrt{\log|T|} \qquad (4)
\end{aligned}
$$

### 3.2 Method 2: $\varepsilon$-net

Let $T' \subseteq T$ be an $\varepsilon$-net of $T$ under $\ell_2$. That is, for all $t \in T$ there exists $t' \in T'$ such that $\|t - t'\|_2 \le \varepsilon$. Now note $\langle g, t\rangle = \langle g, t' + (t - t')\rangle$ so that

$$X_t = X_{t'} + X_{t-t'}.$$

Therefore

$$g(T) \le g(T') + \mathbb{E}\sup_{t\in T}\langle g, t - t'\rangle .$$

We already know $g(T') \lesssim \rho_{\ell_2}(T')\cdot\sqrt{\log|T'|} \le \rho_{\ell_2}(T)\cdot\sqrt{\log|T'|}$ by (4). Also, $\langle g, t - t'\rangle \le \|g\|_2 \cdot \|t - t'\| \le \varepsilon\|g\|_2$, and

$$\mathbb{E}\,\|g\|_2 \le (\mathbb{E}\,\|g\|_2^2)^{1/2} \le \sqrt{n}.$$

Therefore

$$
\begin{aligned}
g(T) &\lesssim \rho_{\ell_2}(T)\cdot\sqrt{\log|T'|} + \varepsilon\sqrt{n} \\
&= \rho_{\ell_2}(T)\cdot\log^{1/2}\mathcal{N}(T, \ell_2, \varepsilon) + \varepsilon\sqrt{n} \qquad (5)
\end{aligned}
$$

where $\mathcal{N}(T, d, u)$ denotes the *entropy number* or *covering number*, defined as the minimum number of radius-$u$ balls under metric $d$ centered at points in $T$ required to cover $T$ (i.e. the size of the smallest $u$-net). Of course $\varepsilon$ can be chosen to minimize (5). Note the case $\varepsilon = 0$ just reduces back to method 1.

### 3.3 Method 3: Dudley's inequality (chaining)

The idea of Dudley's inequality [13] is to, rather than use one net, use a countably infinite sequence of nets. That is, let $S_r \subset T$ denote an $\varepsilon_r$-net of $T$ under $\ell_2$, where $\varepsilon_r = 2^{-r} \cdot \rho_{\ell_2}(T)$. Let $t_r$ denote the closest point in $S_r$ to some $t \in T$. Note $T_0 = \{0\}$ is a valid $\varepsilon_0$-net. Then

$$\langle g, t \rangle = \langle g, t_0 \rangle + \sum_{r=1}^{\infty} \langle g, t_r - t_{r-1} \rangle ,$$

so then

$$\begin{aligned}
g(T) &\leq \sum_{r=1}^{\infty} \mathbb{E} \sup_{t \in T} \langle g, t_r - t_{r-1} \rangle \\
&\lesssim \sum_{r=1}^{\infty} \frac{\rho_{\ell_2}(T)}{2^r} \cdot \log^{1/2}(\mathcal{N}(T, \ell_2, \frac{\rho_{\ell_2}(T)}{2^r})^2) \text{ (by (4))} \qquad (6) \\
&\lesssim \sum_{r=1}^{\infty} \frac{\rho_{\ell_2}(T)}{2^r} \cdot \log^{1/2} \mathcal{N}(T, \ell_2, \frac{\rho_{\ell_2}(T)}{2^r}) \qquad (7)
\end{aligned}$$

where (6) used the triangle inequality to yield

$$\|t_r - t_{r-1}\|_2 \leq \|t - t_r\|_2 + \|t - t_{r-1}\|_2 \leq \frac{3}{2^r} \cdot \rho_{\ell_2}(T).$$

The sum (7) is perfectly fine as is, though the typical formulation of Dudley's inequality then bounds the sum by an integral over $\varepsilon$ (representing $\rho_{\ell_2}(T)/2^r$) then performs the change of variable $u = \varepsilon/\rho_{\ell_2}(T)$. This yields the usual formulation of Dudley's inequality:

$$g(T) \lesssim \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du \qquad (8)$$

It is worth pointing out that Dudley's inequality is equivalent to the following bound. We say $T_0 \subset T_1 \subset \ldots \subset T$ is an *admissible sequence* if $|T_0| = 1$ and $|T_r| \leq 2^{2^r}$. Then Dudley's inequality is equivalent to the bound

$$g(T) \lesssim \sum_{r=0}^{\infty} 2^{r/2} \cdot \sup_{t \in T} d_{\ell_2}(t, T_r). \qquad (9)$$

To see this most easily, compare with the bound (7). Note that to minimize $\sup_{t \in T} d_{\ell_2}(t, T_r)$, we should pick the best quality net we can using $2^{2^r}$ points. From $r = 0$ until some $r_1$, the quality of the net will be, up to a factor of 2, equal to

$\rho_{\ell_2}(T)$, and for the $r$ in this range the summands of (9) will be a geometric series that sum to $O(2^{r_1/2} \cdot \rho_{\ell_2}(T))$. Then from $r = r_1$ to some $r_2$, the quality of the best net will be, up to a factor of 2, equal to $\rho_{\ell_2}(T)/2$, and these summands then are a geometric series that sum to $O(2^{r_2/2} \cdot \rho_{\ell_2}(T)/2)$, etc. In this way, the bounds of (7) and (9) are equivalent up to a constant factor.

Note, this is the primary reason we chose the $T_r$ to have doubly exponential size in $r$: so that the sum of $\log^{1/2}|T_r|$ in any contiguous range of $r$ is a geometric series dominated by the last term.

### 3.4 Method 4: generic chaining

Here we will show the generic chaining method, which yields the bound of [14], though we will present an equivalent bound that was later given by Talagrand (see his book [29]):

$$g(T) \lesssim \inf_{\{T_r\}_{r=0}^{\infty}} \sup_{t\in T} \sum_{r=0}^{\infty} 2^{r/2} \cdot d_{\ell_2}(t, T_r). \tag{10}$$

where the infimum is taken over admissible sequences.

Note the similarity between (9) and (10): the latter bound moved the supremum *outside* the sum. Thus clearly the bound (10) can only be a tighter bound. For a metric $d$, Talagrand defined

$$\gamma_p(T, d) := \inf_{\{T_r\}_r} \sup_{t\in T} \sum_{r=0}^{\infty} 2^{r/p} \cdot d(t, T_r),$$

where again the infimum is over admissible sequences. We now we wish to prove

$$g(T) \lesssim \gamma_2(T, \ell_2).$$

You are probably guessing at this point that had we not been working with subgaussians, but rather random variables that have decay bounded by $e^{-|x|^p}$, we would get a bound in terms of the $\gamma_p$-functional — your guess is right. I leave it to you as an exercise to modify arguments appropriately!

For nonnegative integer $r$ and for $t \in T$, define $\pi_r t = \mathrm{argmin}_{t'\in T_r} d(t, t')$. For $r \geq 1$ define $\Delta_r t = \pi_r t - \pi_{r-1} t$. Then for any $t \in T$

$$t = \pi_0 t + \sum_{r=1}^{\infty} \Delta_r t$$

so that

$$\mathbb{E} \sup_{t\in T} \langle g, t\rangle = \mathbb{E} \sup_{t\in T} \sum_{r=1}^{\infty} \underbrace{\langle g, \Delta_r t\rangle}_{Y_r(t)}.$$

since $\mathbb{E}\, sup_{t\in T} \langle g, \pi_0 t\rangle = \mathbb{E}\langle g, \pi_0 t\rangle = 0$, with the first equality using that $|T_0| = 1$.

Note for fixed $t$, by gaussian decay

$$\mathbb{P}(|Y_r(t)| > 2u2^{r/2}\|\Delta_r t\|) < 2e^{-u^2 2^r}.$$

Therefore

$$\begin{aligned}
\mathbb{P}(\exists t \in T, r > 0 \ s.t.\ |Y_r(t)| > 2u2^{r/2}\|\Delta_r t\|) &\lesssim \sum_{r=1}^{\infty} |T_r| \cdot |T_{r-1}| \cdot e^{-u^2 2^r} \\
&\leq \sum_{r=1}^{\infty} 4^{2^r} \cdot e^{-u^2 2^r} \qquad (11)
\end{aligned}$$

since $|T_r|, |T_{r-1}| \leq 2^{2^r}$. The above sum is convergent for $u \geq 2$.

Now, again using that $\mathbb{E}\,|Z| = \int_0^\infty \mathbb{P}(|Z| > w)dw$, we have

$$\begin{aligned}
g(T) &\leq \int_0^\infty \mathbb{P}(\sup_{t\in T} \sum_{r=1}^{\infty} Y_r > w)dw \\
&= \left(2 \sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|\right) \\
&\quad \times \int_0^\infty \mathbb{P}(\sup_{t\in T} \sum_{r=1}^{\infty} Y_r > u \cdot 2 \sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|)du \text{ (change of variables)} \\
&\lesssim \left(\sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|\right) \\
&\quad \times [2 + \int_2^\infty \mathbb{P}(\sup_{t\in T} \sum_{r=1}^{\infty} Y_r > u \cdot 2 \sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|)du] \\
&\lesssim \left(\sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|\right) \\
&\quad \times [2 + \int_2^\infty \mathbb{P}(\exists t \in T, r > 0 \ s.t.\ |Y_r(t)| > 2u2^{r/2}\|\Delta_r t\|)du] \\
&\lesssim \sup_{t\in T} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r t\|. \qquad (12)
\end{aligned}$$

Now note $\|\Delta_r t\| = \|t_r - t_{r-1}\| \leq 2d_{\ell_2}(t, T_r)$ by the triangle inequality, and thus (12) is at most a constant factor larger than $\gamma_2(T, \ell_2)$, as desired.

Surprisingly, Talagrand showed that not only is $\gamma_2(T, \ell_2)$ an asymptotic upper bound for $g(T)$, but it is also an asymptotic *lower bound* (at least when the entries of $g$ are *actually* gaussians — the lower bound does not hold for subgaussian entries). That is, $g(T) \simeq \gamma_2(T, \ell_2)$ for *any* $T$. This is known as the "majorizing measures theorem" for reasons we will not get into. In brief: the formulation of [14] did not talk about admissible sequences, or discrete sets at all, but rather worked with measures and provided an upper bound in terms of an infimum over a set of probability measures of a certain integral — this formulation is equivalent to the formulation discussed above in terms of admissible sets, and a proof of the equivalence appears in [29].

# 4 A concrete example: the $\ell_1$ ball

Consider the example $T = B_{\ell_1^n} = \{t \in \mathbb{R}^n : \|t\|_1 = 1\}$, i.e. the unit $\ell_1$. I picked this example because it is easy to already know $g(T)$ using other methods. Why? Well, $\sup_{t \in B_{\ell_1^n}} \langle g, t\rangle = \|g\|_\infty$, since the dual norm of $\ell_\infty$ is $\ell_1$! Thus $g(B_{\ell_1^n}) = \mathbb{E}\,\|g\|_\infty$, which one can check is $\Theta(\sqrt{\log n})$. Thus we *know* the answer is $\Theta(\sqrt{\log n})$.

So now the question: what do the four methods above give?

## 4.1 Method 1: union bound

This method gives nothing, since $T$ is an infinite set.

## 4.2 Method 2: $\varepsilon$-net

To apply this method, we need to understand the size of an $\varepsilon$-net of the $\ell_1$ unit ball under $\ell_2$. One bound comes from Maurey's empirical method.

**Lemma 1** (Maurey's empirical method)**.** $\mathcal{N}(B_{\ell_1^n}, \ell_2, u) \le (2n)^{4/u^2}$

*Proof.* Consider any $t \in B_{\ell_1^n}$. It can be written as a convex combination $t = \sum_{i=1}^{2n} \alpha_i x_i$ where $x_1, \ldots, x_n = e_1, \ldots, e_n$ and $x_{n+1}, \ldots, x_{2n} = -e_1, \ldots, -e_n$. Now, consider a distribution over $\mathbb{R}^n$ in which we pick a random vector $v$ which equals $t_i$ with probability $\alpha_i$. Then $\mathbb{E}\, v = t$. Now pick $Z_1, \ldots, Z_q, Z_1', \ldots, Z_q'$ i.i.d. from this distribution. Define the vectors $Z = (Z_1, \ldots, Z_q)$ and $Z' = (Z'_1, \ldots, Z'_q)$. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z} \|t - \frac{1}{q}\sum_{i=1}^{q} Z_i\|_2 &= \frac{1}{q}\mathop{\mathbb{E}}_{Z} \| \mathop{\mathbb{E}}_{Z'} \sum_{i=1}^{q} (Z_i - Z'_i)\|_2 \\
&= \frac{1}{q}\mathop{\mathbb{E}}_{Z} \| \mathop{\mathbb{E}}_{\sigma, Z'} \sum_{i=1}^{q} \sigma_i (Z_i - Z'_i)\|_2 \\
&\le \frac{1}{q}\mathop{\mathbb{E}}_{Z, Z', \sigma} \| \sum_{i=1}^{q} \sigma_i (Z_i - Z'_i)\|_2 \text{ (Jensen)} \\
&\le \frac{2}{q}\mathop{\mathbb{E}}_{Z}\mathop{\mathbb{E}}_{\sigma} \| \sum_{i=1}^{q} \sigma_i Z_i\|_2 \\
&\le \frac{2}{q}\mathop{\mathbb{E}}_{Z}(\mathop{\mathbb{E}}_{\sigma} \| \sum_{i=1}^{q} \sigma_i Z_i\|_2^2)^{1/2} \\
&= \frac{2}{\sqrt{q}}.
\end{aligned}$$

where the $\sigma_i$ are independent uniform $\pm 1$ random variables. Thus, in expectation, $t$ is $u$-close to an average of $q$ such random $Z_i$ for $q \ge 4/u^2$. Thus in particular, every $t$ in $B_{\ell_1^n}$ is $u$-close in $\ell_2$ to *some* average of $4/u^2$ of the vectors $\pm e_i$, and thus the set of all such averages is a $u$-net in $\ell_2$, of which there are at most $(2n)^q$. □

One can also obtain a bound on the covering number via a simple volumetric argument, which implies $\mathcal{N}(B^n_{\ell_1}, \ell_2, \varepsilon) = O(2 + 1/(u\sqrt{n}))^n$. Without giving the precise calculations, the argument is to first *upper bound* the maximum number of disjoint radius $(u/2)$-$\ell_2$ balls one can pack in $B_{\ell_1^n}$. Then if one takes those balls and considers the union of radius-$u$ balls from their centers, these balls must cover of $B_{\ell_1^n}$ by the triangle inequality and maximality of the original packing. Since all the original packed balls are fully contained in the $\ell_1$ ball of radius $1 + (u/2)\sqrt{n}$ by Cauchy-Schwarz, the number of balls in the packing could not have been more than the ratio of the volume of the $\ell_1$ ball of radius $(1 + (u/2)\sqrt{n})$, and the volume of an $\ell_2$ ball of radius $u/2$. Thus, combining Maurey's lemma and this argument,

$$\forall \varepsilon \in (0, \frac{1}{2}),\ \log^{1/2} \mathcal{N}(B^n_{\ell_1}, \ell_2, \varepsilon) \lesssim \min\{\varepsilon^{-1}\sqrt{\log n},\ \sqrt{n} \cdot \log(1/\varepsilon)\}. \tag{13}$$

By picking $\varepsilon = ((\log n)/n)^{1/4}$, (5) gives us $g(T) \lesssim (n \log n)^{1/4}$. This is exponentially worse than true bound of $g(T) = \Theta(\sqrt{\log n})$.

## 4.3 Method 3: Dudley's inequality

Combining (13) with (8),

$$g(T) \lesssim \int_0^{1/\sqrt{n}} \sqrt{n} \cdot \log(1/u) du + \int_{1/\sqrt{n}}^1 u^{-1} \sqrt{\log n} du \lesssim \log^{3/2} n.$$

This is exponentially better than method 2, but still off from the truth. We can though wonder: perhaps the issue is not Dudley's inequality, but perhaps the entropy bounds of (13) are simply loose? Unfortunately this is not the case. To see this, take a set $R$ of vectors in $\mathbb{R}^n$ that are each $1/\varepsilon^2$-sparse, with $\varepsilon^2$ in each non-zero coordinate, and so that all pairwise $\ell_2$ distances are $2\varepsilon$. A random collection $R$ satisfies this distance property with high probability for $|R| = n^{\Theta(1/\varepsilon^2)}$ and $\varepsilon \gg 1/\sqrt{n}$. Then note $R \subset B_{\ell_1^n}$ and furthermore one needs at least $|R|$ radius-$\varepsilon$ balls in $\ell_2$ just to cover $R$.

It is also worth pointing out that this is the worst case for Dudley's inequality: it can never be off by more than a factor of $\log n$. I'll leave it to you as an exercise to figure out why (you should assume the majorizing measures theorem, i.e. that (10) is tight)! **Hint:** compare (9) with (10) and show that nothing interesting happens beyond $r > \log n + c \log\log n$.

## 4.4 Method 4: generic chaining

By the majorizing measures theorem, we *know* there must exist an admissible sequence giving the correct $g(T) \lesssim \sqrt{\log n}$, thus being superior to Dudley's inequality. Once as an exercise, I tried with Eric Price and Mary Wootters to construct an explicit admissible sequence demonstrating that $\gamma_2(B_{\ell_1^n}, \ell_2) = O(\sqrt{\log n})$. Eric and I managed to find a sequence yielding $O(\log n)$, and Mary found a sequence that gives the correct $O(\sqrt{\log n})$ bound. Below I include Mary's construction.

Henceforth, to be concrete log denotes $\log_2$. Let $\mathcal{N}_s$ be a $1/2^s$-net of the $2^s$-sparse vectors in $B_{\ell_2^n}$. Thus

$$|\mathcal{N}_s| \le \binom{n}{2^s}(3 \cdot 2^s)^{2^s}.$$

Then defining $s_k = k - \lceil \log\log(3en) \rceil$,

$$|\mathcal{N}_{s_k}| \le 2^{2^k}.$$

Then we define $T_0 = T_1 = \cdots = T_{\lceil \log\log(3en)\rceil - 1} = \{0\}$, and $T_k = \mathcal{N}_{s_k}$ for $\lceil \log\log(3en) \rceil \le k \le \ell_{max}$ for $\ell_{max} = \log n + \lceil \log\log(3en) \rceil$. For $k \ge \ell_{max}$, we set $T_k$ to be an $\varepsilon_k$-net of $B_{\ell_2^n}$ of size $2^{2^k}$ for the smallest $\varepsilon_k$ possible. If $k = \ell_{max} + j$, then

$$\varepsilon \le n^{-2^j}.$$

We now wish to upper bound the supremum over all $x \in B_{\ell_1^n}$ of

$$\sum_{k=0}^{\infty} 2^{k/2} d_{\ell_2}(x, T_k). \tag{14}$$

We henceforth focus on a particular $x \in B_{\ell_1^n}$ and show that (14) is $O(\sqrt{\log n})$. We split the sum into three parts:

(1) $0 \le k < \lceil \log\log(3en) \rceil$

(2) $\lceil \log\log(3en) \rceil \le k < \ell_{max}$

(3) $\ell_{max} \le k < \infty$

For the summands in (1), each $d_{\ell_2}(x, T_k)$ equals $\|x\|_2 \le 1$, and thus these terms in total contribute at most $2 \cdot 2^{\lceil \log\log(3en) \rceil} = O(\sqrt{\log n})$ to (14). The summands in (3) are also easy to handle: writing $k = \ell_{max} + j$, the summand with index $k$ is at most

$$2^{(\ell_{max}+j)/2} \cdot n^{-2^j} \le \sqrt{n \log n} \cdot 2^{j/2} n^{-2^j},$$

and thus the sum over $j \ge 0$ is $o(1)$ for any $n \ge 2$.

We now proceed with the most involved part of the argument: bounding the contribution of summands in the range (2). For this, we will use a technique that is often referred to in the compressed sensing community as *shelling*. Consider sorting the indices $i \in [n]$ by magnitude $|x_i|$, i.e. $|x_{i_1}| \ge |x_{i_2}| \ge \ldots \ge |x_{i_n}|$. Define the vector $|x|$ by $|x|_i = |x_i|$. Let $A_0 \subset [n]$ denote the coordinates of the $2^0$ largest entries of $|x|$, then $A_1$ the next $2^1$ largest entries, then $A_2$ the next $2^2$ largest entries, etc. (if less than $2^s$ entries remain in $x$, then $A_s$ is simply the set of remaining entries). The $A_s$ partition $[n]$. Let $x_A \in \mathbb{R}^n$ denote the projection of $x$ onto coordinates in $A$.

$$\begin{aligned}
\sum_{k=\lceil \log\log(3en) \rceil}^{\log n + \lceil \log\log(3en) \rceil} 2^{k/2} \cdot d_{\ell_2}(x, \mathcal{N}_{s_k}) &\lesssim \sqrt{\log n} \cdot \sum_{s=0}^{\log n} 2^{s/2} \cdot d_{\ell_2}(x, \mathcal{N}_s) \\
&\lesssim \sqrt{\log n} \cdot \sum_{s=0}^{\log n} 2^{s/2} \cdot \left( d_{\ell_2^n}(x_{A_s}, \mathcal{N}_s) + \|x - x_{A_s}\|_2 \right) \\
&\lesssim \sqrt{\log n} + \sqrt{\log n} \cdot \underbrace{\sum_{s=0}^{\log n} 2^{s/2} \cdot \|x - x_{A_s}\|_2}_{\alpha}
\end{aligned}$$

We now wish to show $\alpha = O(1)$.

$$\begin{aligned}\alpha &\le \sum_{s=0}^{\log n} 2^{s/2}\left(\sum_{j=s+1}^{\log n} \|x_{A_j}\|_2\right)\\ &\le \sum_{s=0}^{\log n} 2^{s/2}\cdot\left(\sum_{j=s+1}^{\log n} 2^{j/2}\|x_{A_j}\|_\infty\right)\\ &= \sum_{j=1}^{\log n} 2^{j/2}\|x_{A_j}\|_\infty\cdot\left(\sum_{s=0}^{j-1} 2^{s/2}\right)\\ &\lesssim \sum_{j=1}^{\log n} 2^{j}\cdot\|x_{A_j}\|_\infty \qquad (15)\end{aligned}$$

The largest entry of $|x|_{A_j}$ is at most the smallest entry of $|x|_{A_{j-1}}$ by construction, and hence is at most the average entry of $|x|_{A_{j-1}}$. Thus

$$\begin{aligned}(15) &\le \sum_{j=1}^{\log n} 2^{j}\cdot\frac{\|x_{A_{j-1}}\|_1}{2^{j-1}}\\ &\le 2\cdot\sum_{j=0}^{\log n-1}\|x_{A_j}\|_1\\ &\le 2\cdot\|x\|_1,\end{aligned}$$

which is at most $2 = O(1)$, as desired.

# 5 Application details: dimensionality reduction

We again use the definitions of $\pi_r, \Delta_r$ from Section 3.4. Also, throughout this section we let $\|\cdot\|$ denote the $\ell_{2\to 2}$ operator norm in the case of matrix arguments, and the $\ell_2$ norm in the case of vector arguments. Recall $\rho_X(T)$ denotes diameter of $T$ under norm $\|\cdot\|_X$. We use $\|\cdot\|_F$ to denote Frobenius norm.

Krahmer, Mendelson, and Rauhut showed the following theorem [20].

**Theorem 1.** *Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$ be arbitrary. Let $\sigma_1, \ldots, \sigma_n$ be independent subgaussian random variables of mean* 0 *and variance* 1. *Then*

$$\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}}\left|\|A\sigma\|^2 - \mathbb{E}\,\|A\sigma\|^2\right| \lesssim \gamma_2^2(\mathcal{A}, \|\cdot\|) + \gamma_2(\mathcal{A}, \|\cdot\|)\cdot\rho_F(\mathcal{A}) + \rho_F(\mathcal{A})\cdot\rho_{\ell_{2\to 2}}(\mathcal{A}).$$

We now show that Theorem 1, combined with the majorizing measures theorem, can be used to prove the theorem of Gordon [15] as described in Section 2, and in fact a theorem that is slightly stronger. Gordon's original proof did not use chaining at all. Recall from (2) that we have a point set $P \subset \mathbb{R}^d$, and we want to show that a random matrix $\Pi \in \mathbb{R}^{m\times n}$ satisfies

$$\forall x, y \in P,\ (1-\varepsilon)\|x-y\|_2^2 \le \|\Pi x - \Pi y\|_2^2 \le (1+\varepsilon)\|x-y\|_2^2.$$

for $m$ not too large. In other words, for $T = \{(x-y)/\|x-y\| : x \neq y \in P\}$, we want

$$\sup_{x\in T} \left| \|\Pi x\|^2 - 1 \right| < \varepsilon. \tag{16}$$

We below show that Theorem 1 implies that the *expectation* of the left hand side of (16) is less than $\varepsilon$ for $m \gtrsim (g^2(T)+1)/\varepsilon^2$, when the entries of $\Pi$ are i.i.d. subgaussian with mean 0 and variance $1/m$. Gordon showed the same result but only when the $\Pi_{i,j}$ were independent gaussians and not subgaussians. Note bounding the expectation by $\varepsilon$ in (16) implies the actual sup is at most $3\varepsilon$ with probability 2/3, by Markov's inequality. Much stronger concentration analyses have been given by bounding the $L^p$ norm of the left hand side then performing Markov's inequality on a high moment [24, 9, 10]; we do not cover those approaches here.

We only show the theorem when $T$ is finite. In many applications we care about infinite $T$ (e.g. all the unit norm vectors in a $d$-dimensional subspace, for applications in numerical linear algebra [31]). In fact, for $T \subset \ell_2^n$ bounded it is without loss of generality to consider only finite $T$. This is because we can take $T'$ a finite $\alpha$-net of $T$, i.e. $\forall x \in T\ \exists x' \in T' : \|x - x'\| \le \alpha$. Then

$$g(T) = \mathbb{E}_g \sup_{x\in T} \langle g, x'\rangle + \langle g, x - x'\rangle = g(T') \pm \mathbb{E}_g \sup_{x\in T} \langle g, x' - x\rangle = g(T') \pm \alpha\sqrt{n}$$

since $|\langle g, x' - x\rangle| \le \|g\| \cdot \|x - x'\|$ and $\mathbb{E}_g \|g\| \le (\mathbb{E}_g \|g\|^2)^{1/2} = \sqrt{n}$. Then we can choose $\alpha$ arbitrarily small so that $g(T')$ is as close to $g(T)$ as we want.

**Theorem 2.** *Let $T \subset \mathbb{R}^n$ be a finite set of vectors each of unit norm, and let $\varepsilon \in (0, 1/2)$ be arbitrary. Let $\Pi \in \mathbb{R}^{m\times n}$ be such that $\Pi_{i,j} = \sigma_{i,j}/\sqrt{m}$ for independent subgaussian variables $\sigma_{i,j}$ of mean 0 and variance 1, where $m \gtrsim (g^2(T)+1)/\varepsilon^2$. Then*

$$\mathbb{E}_\sigma \sup_{x\in T} \left| \|\Pi x\|^2 - 1 \right| < \varepsilon.$$

*Proof.* For $x \in T$ let $A_x$ denote the $m \times mn$ matrix defined as follows:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

Then $\|\Pi x\|^2 = \|A_x\sigma\|^2$, so letting $\mathcal{A} = \{A_x : x \in T\}$,

$$\mathop{\mathbb{E}}_{\sigma} \sup_{x\in T} \left| \|\Pi x\|^2 - 1 \right| = \mathop{\mathbb{E}}_{\sigma} \sup_{A\in\mathcal{A}} \left| \|A\sigma\|^2 - \mathop{\mathbb{E}}_{\sigma} \|A\sigma\|^2 \right|.$$

We have $\rho_F(\mathcal{A}) = 1$. Also $A_x^* A_x$ is a block-diagonal matrix, with $m$ blocks each equal to $xx^*/m$, and thus the singular values of $A_x$ are 0 and $\|x\|/\sqrt{m}$, implying $\rho_{\ell_{2\to 2}}(\mathcal{A}) = 1/\sqrt{m}$. Similarly, since $A_x - A_y = A_{x-y}$, for any vectors $x, y$ we have $\|A_x - A_y\| = \|x - y\|$, and thus $\gamma_2(\mathcal{A}, \|\cdot\|) \le \gamma_2(T, \|\cdot\|)/\sqrt{m}$. Thus by Theorem 1,

$$\mathop{\mathbb{E}}_{\sigma} \sup_{x\in T} \left| \|\Pi x\|^2 - 1 \right| \lesssim \frac{\gamma_2^2(T, \|\cdot\|)}{m} + \frac{\gamma_2(T, \|\cdot\|)}{\sqrt{m}} + \frac{1}{\sqrt{m}},$$

which is at most $\varepsilon$ for $m \gtrsim (\gamma_2^2(T, \|\cdot\|) + 1)/\varepsilon^2$ as in the theorem statement. This inequality holds by setting $m \gtrsim (g^2(T) + 1)/\varepsilon^2$, since $\gamma_2(T, \|\cdot\|) \lesssim g(T)$ by the majorizing measures theorem. $\square$

We now prove Theorem 1. We only prove it in the case that the $\sigma_i$ are Rademacher, i.e. uniform $\pm 1$, since this setting already contains the main ideas of the proof. Before we can continue with the proof though, we need a few standard lemmas. The proofs given below are also standard. Recall that for a scalar random variable $Z$, $\|Z\|_p$ denotes $(\mathbb{E}\,|Z|^p)^{1/p}$. It is known that $\|\cdot\|_p$ is a norm for $p \ge 1$.

**Lemma 2** (Khintchine's inequality)**.** *Let $x \in \mathbb{R}^n$ be arbitrary and $\sigma_1, \ldots, \sigma_n$ be independent Rademachers. Then*

$$\forall p \ge 1,\ \|\langle \sigma, x\rangle\|_p \le \sqrt{p} \cdot \|x\|.$$

*This is equivalent, up to constant factors in the exponent, to the following:*

$$\forall \lambda > 0,\ \mathop{\mathbb{P}}_{\sigma}(|\langle \sigma, x\rangle| > \lambda) \le 2e^{-\lambda^2/(2\|x\|^2)}.$$

*Proof.* For the first inequality, consider $\langle g, x\rangle$ for $g$ a vector of independent standard normal random variables. The random variable $\langle g, x\rangle$ is distributed as a gaussian with variance $\|x\|^2$, and thus $\|\langle g, x\rangle\|_p < \sqrt{p} \cdot \|x\|$ by known moment bounds on gaussians. Meanwhile, for positive even integer $p$, one can expand $\mathbb{E}\,|\langle g, x\rangle^p| = \mathbb{E}\,\langle g, x\rangle^p$ as a sum of expectations of monomials. If one similarly expands $\langle \sigma, x\rangle^p$, then we find that these monomials' expectations are term-by-term dominated in the gaussian case, since any even Rademacher moment is 1 whereas all even gaussian moments are at least 1. $\square$

**Lemma 3** (Decoupling [12]). *Let $x_1, \ldots, x_n$ be independent and mean zero, and $x'_1, \ldots, x'_n$ identically distributed as the $x_i$ and independent of them. Then for any $(a_{i,j})$ and for all $p \geq 1$*

$$\|\sum_{i \neq j} a_{i,j} x_i x_j\|_p \leq 4 \|\sum_{i,j} a_{i,j} x_i x'_j\|_p$$

*Proof.* Let $\eta_1, \ldots, \eta_n$ be independent Bernoulli random variables each of expectation $1/2$. Then

$$\begin{aligned} \|\sum_{i \neq j} a_{i,j} x_i x_j\|_p &= 4 \cdot \|\mathop{\mathbb{E}}_{\eta} \sum_{i \neq j} a_{i,j} x_i x_j |\eta_i||1 - \eta_j|\|_p \\ &\leq 4 \cdot \|\sum_{i \neq j} a_{i,j} x_i x_j \eta_i (1 - \eta_j)\|_p \text{ (Jensen)} \end{aligned} \tag{17}$$

Hence there must be some fixed vector $\eta' \in \{0, 1\}^n$ which achieves

$$\|\sum_{i \neq j} a_{i,j} x_i x_j \eta_i (1 - \eta_j)\|_p \leq \|\sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x_j\|_p$$

where $S = \{i : \eta'_i = 1\}$. Let $x_S$ denote the $|S|$-dimensional vector corresponding to the $x_i$ for $i \in S$. Then

$$\begin{aligned} \|\sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x_j\|_p &= \|\sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x'_j\|_p \\ &= \|\mathop{\mathbb{E}}_{x_S} \mathop{\mathbb{E}}_{x'_{\bar{S}}} \sum_{i,j} a_{i,j} x_i x'_j\|_p \text{ } (\mathbb{E}\, x_i = \mathbb{E}\, x'_j = 0) \\ &\leq \|\sum_{i,j} a_{i,j} x_i x'_j\|_p \text{ (Jensen)} \end{aligned}$$

□

## 5.1 Proof of Theorem 1

We now prove Theorem 1 in the case the $\sigma_i$ are independent Rademachers. Without loss of generality we can assume $\mathcal{A}$ is finite (else apply the theorem to a sufficiently fine net, i.e. fine in $\ell_2 \to \ell_2$ operator norm). Define

$$E = \mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}} \left| \|A\sigma\|^2 - \mathbb{E}\, \|A\sigma\|^2 \right|$$

and let $A^i$ denote the $i$th column of $A$. Then by decoupling

$$E = \mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}} \left| \sum_{i \neq j} \sigma_i \sigma_j \left\langle A^i, A^j \right\rangle \right|$$

$$\leq 4 \cdot \mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} \left| \sum_{i,j} \sigma_i \sigma'_j \left\langle A^i, A^j \right\rangle \right|$$
$$= 4 \cdot \mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} |\langle A\sigma, A\sigma'\rangle| .$$

Let $\{T_r\}_{r=0}^{\infty}$ be admissible for $\mathcal{A}$. Direct computation shows

$$\langle A\sigma, A\sigma'\rangle = \langle(\pi_0 A)\sigma, (\pi_0 A)\sigma'\rangle + \sum_{r=1}^{\infty} \underbrace{\langle(\Delta_r A)\sigma, (\pi_{r-1}A)\sigma'\rangle}_{X_r(A)} + \sum_{r=1}^{\infty} \underbrace{\langle(\pi_r A)\sigma, (\Delta_r A)\sigma'\rangle}_{Y_r(A)} .$$

We have $T_0 = \{A_0\}$ for some $A_0 \in \mathcal{A}$. Thus $\mathbb{E}_{\sigma,\sigma'} |\langle(\pi_0 A)\sigma, (\pi_0 A)\sigma'\rangle|$ equals

$$\mathop{\mathbb{E}}_{\sigma,\sigma'} \left|\sigma^* A_0^* A_0 \sigma'\right| \leq \left(\mathop{\mathbb{E}}_{\sigma,\sigma'} (\sigma^* A_0^* A_0 \sigma')^2\right)^{1/2} = \|A_0^* A_0\|_F \leq \|A_0\|_F \|A_0\| \leq \rho_F(\mathcal{A}) \cdot \rho_{\ell_{2\to 2}}(\mathcal{A}).$$

Thus,

$$\mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} |\langle A\sigma, A\sigma'\rangle| \leq \rho_F(\mathcal{A}) \cdot \rho_{\ell_{2\to 2}}(\mathcal{A}) + \mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} |X_r(A)| + \mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} |Y_r(A)|.$$

We focus on the second summand; handling the third summand is similar.

Note $X_r(A) = \langle(\Delta_r A)\sigma, (\pi_{r-1}A)\sigma'\rangle = \langle\sigma, (\Delta_r A)^*(\pi_{r-1}A)\sigma'\rangle$. Thus by the Khintchine inequality (namely $\| \langle\sigma, x\rangle \|_p \lesssim \sqrt{p} \cdot \|x\|$),

$$\mathbb{P}(|X_r(A)| > t2^{r/2} \cdot \|(\Delta_r A)^*(\pi_{r-1}A)\sigma'\|) \lesssim e^{-t^2 2^r/2}.$$

Let $\mathcal{E}(A)$ be the event that for all $r \geq 1$ simultaneously, $|X_r(A)| \leq t2^{r/2} \cdot \|\Delta_r A\| \cdot \sup_{A\in\mathcal{A}} \|A\sigma'\|$. Then

$$\mathbb{P}(\exists A \in \mathcal{A}\ s.t.\ \neg\mathcal{E}(A)) \lesssim \sum_{r=1}^{\infty} |T_r| \cdot |T_{r-1}| \cdot e^{-t^2 2^r/2}$$
$$\leq \sum_{r=1}^{\infty} 2^{2^{r+1}} \cdot e^{-t^2 2^r/2}.$$

Therefore

$$\mathop{\mathbb{E}}_{\sigma,\sigma'} \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} |X_r(A)| = \mathop{\mathbb{E}}_{\sigma'} \int_0^{\infty} \mathop{\mathbb{P}}_{\sigma} \left( \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} |X_r(A)| > t \right) dt,$$

which by a change of variables is equal to

$$\mathop{\mathbb{E}}_{\sigma'} \left( \sup_{A\in\mathcal{A}} \|A\sigma'\| \cdot \left( \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} 2^{r/2} \|\Delta_r A\| \right) \right.$$

$$\times \cdot \int_0^\infty \mathbb{P}_{\sigma}\left(\sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} |X_r(A)| > t \sup_{A\in\mathcal{A}} 2^{r/2} \cdot \|\Delta_r A\| \cdot \sup_{A\in\mathcal{A}} \|A\sigma'\|\right) dt\Bigg)$$

$$\leq \left(\mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \|A\sigma'\|\right) \cdot \left(\sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r A\|\right) \cdot \left[3 + \sum_{r=1}^{\infty} \int_3^\infty 2^{2^{r+1}} e^{-t^2 2^r/2} dt\right]$$

$$\lesssim \left(\mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \|A\sigma'\|\right) \cdot \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} 2^{r/2}\|\Delta_r A\|$$

$$\lesssim \left(\mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \|A\sigma'\|\right) \cdot \sup_{A\in\mathcal{A}} \sum_{r=1}^{\infty} 2^{r/2} \cdot \rho_{2\to 2}(A, T_r),$$

since $\|\Delta_r A\| \leq \rho_{2\to 2}(A, T_{r-1}) + \rho_{2\to 2}(A, T_r)$ via the triangle inequality. Choosing admissible $T_0 \subseteq T_1 \subseteq \ldots \subseteq T$ to minimize the above expression,

$$E \lesssim \rho_F(\mathcal{A}) \cdot \rho_{\ell_{2\to 2}}(\mathcal{A}) + \gamma_2(\mathcal{A}, \|\cdot\|) \cdot \mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \|A\sigma'\|.$$

Now observe

$$\begin{aligned}
\mathbb{E}_{\sigma'}\left(\sup_{A\in\mathcal{A}} \|A\sigma'\|\right) &\leq \left(\mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \|A\sigma'\|^2\right)^{1/2} \\
&\leq \left(\mathbb{E}_{\sigma'}\left(\sup_{A\in\mathcal{A}} \left| \|A\sigma'\|^2 - \mathbb{E}_{\sigma'} \|A\sigma'\|^2 \right| + \mathbb{E}_{\sigma'} \|A\sigma'\|^2\right)\right)^{1/2} \\
&= \left(\mathbb{E}_{\sigma'} \sup_{A\in\mathcal{A}} \left(\left| \|A\sigma'\|^2 - \mathbb{E}_{\sigma'} \|A\sigma'\|^2 \right| + \|A\|_F^2\right)\right)^{1/2} \\
&\leq \sqrt{E} + \rho_F(\mathcal{A})
\end{aligned}$$

Thus in summary,

$$E \lesssim \rho_F(\mathcal{A}) \cdot \rho_{\ell_{2\to 2}}(\mathcal{A}) + \gamma_2(\mathcal{A}, \|\cdot\|) \cdot (\sqrt{E} + \rho_F(\mathcal{A})).$$

This implies $E$ is at most the square of the larger root of the associated quadratic equation, which gives the theorem.

# Acknowledgments

I thank Oded Regev for pointing out that Dudley's inequality cannot yield a bound on the mean width of $B_{\ell_1^n}$ that is better than $O(\log^{3/2} n)$ (see Section 4.3), and I thank Mary Wootters for allowing me to include her argument for the admissible sequence for the $\ell_1$ ball in Section 4. A preliminary version of this note appeared on the "Windows on Theory" blog. I thank commenters there, in addition to Assaf Naor and David Woodruff, for pointing out typographical errors and for other suggestions.

# References


[1] Radosław Adamczak. A note on the sample complexity of the Er-SpUD algorithm by Spielman, Wang and Wright for exact recovery of sparsely used dictionaries. *CoRR*, abs/1601.02049, 2016.

[2] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013.

[3] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[4] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P. Woodruff. BPTree: an $\ell_2$ heavy hitters algorithm using constant memory. *CoRR*, abs/1603.00759, 2016.

[5] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, and David P. Woodruff. Beating CountSketch for Heavy Hitters in Insertion Streams. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 740–753, 2016. Full version at arXiv abs/1511.00661.

[6] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geometric and Functional Analysis (GAFA)*, 25(4):1009–1088, July 2015. Preliminary version in STOC 2015.

[7] Jaroslaw Blasiok and Jelani Nelson. An improved analysis of the er-spud dictionary learning algorithm. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 44:1–44:14, 2016. Full version at arXiv abs/1602.05719.

[8] Jian Ding. Asymptotics of cover times via Gaussian free fields: Bounded-degree graphs and general trees. *Annals of Probability*, 42(2):464–496, 2014.

[9] Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *Found. Comput. Math.*, pages 1–30, 2015. Full version at arXiv abs/1402.3973.

[10] Sjoerd Dirksen. Tail bounds via generic chaining. *Electron. J. Probab.*, 20(53):1–29, 2015.

[11] Jian Ding, James R. Lee, and Yuval Peres. Cover times, blanket times, and majorizing measures. *Annals of Mathematics*, 175:1409–1471, 2012.

[12] Victor de la Peña and Evarist Giné. *Decoupling: From dependence to independence*. Probability and its Applications. Springer-Verlag, New York, 1999.

[13] Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[14] Xavier Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. *Lecture Notes in Math.*, 480:1–96, 1975.

[15] Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

[16] Zengfeng Huang, Wai Ming Tai, and Ke Yi. Tracking the frequency moments at all times. *CoRR*, abs/1412.1763, 2014.

[17] Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms*, 3(3), 2007.

[18] T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26, 2013.

[19] Bo'az Klartag and Shahar Mendelson. Empirical processes and random projections. *J. Funct. Anal.*, 225(1):229–245, 2005.

[20] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 2014.

[21] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, 1991.

[22] Kyle Luh and Van Vu. Random matrices: $l_1$ concentration and dictionary learning with few samples. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1409–1425, 2015.

[23] Raghu Meka. A PTAS for computing the supremum of gaussian processes. In *53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 217–222, 2012.

[24] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17:1248–1282, 2007.

[25] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Isometric sketching of any set via the restricted isometry property. *CoRR*, abs/1506.03521, 2015.

[26] Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC)*, pages 764–773, 2014.

[27] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 37.1–37.18, 2012.

[28] Michel Talagrand. Are many small sets explicitly small? In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 13–36, 2010.

[29] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Springer, 2014.

[30] Ramon van Handel. Probability in high dimensions. Manuscript, 2014. Available at `https://www.princeton.edu/~rvan/ORF570.pdf`. Version from June 30, 2014.

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

[32] Alex Zhai. Exponential concentration of cover times. *CoRR*, abs/1407.7617, 2014.