5 Johnson-Lindenstrauss Lemma and Gordons Theorem

5.1 The Johnson-Lindenstrauss Lemma

Suppose one has n points, $X = \{x_1, \ldots, x_n\}$, in $\mathbb{R}^d$ (with d large). If $d > n$, since the points have to lie in a subspace of dimension n it is clear that one can consider the projection $f : \mathbb{R}^d \to \mathbb{R}^n$ of the points to that subspace without distorting the geometry of X. In particular, for every x_i and x_j, $\|f(x_i) - f(x_j)\|^2 = \|x_i - x_j\|^2$, meaning that f is an isometry in X.

Suppose now we allow a bit of distortion, and look for $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry, meaning that

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2. \tag{48}$$

Can we do better than $k = n$?

In 1984, Johnson and Lindenstrauss [JL84] showed a remarkable Lemma (below) that answers this question positively.

Theorem 5.1 (Johnson-Lindenstrauss Lemma [JL84]) *For any $0 < \epsilon < 1$ and for any integer n, let k be such that*

$$k \geq 4\frac{1}{\epsilon^2/2 - \epsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, there is a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry for X (see (48)). This map can be found in randomized polynomial time.

We borrow, from [DG02], an elementary proof for the Theorem. We need a few concentration of measure bounds, we will omit the proof of those but they are available in [DG02] and are essentially the same ideas as those used to show Hoeffding's inequality.

Lemma 5.2 (see [DG02]) *Let $y_1, \ldots, y_d$ be i.i.d standard Gaussian random variables and $Y = (y_1, \ldots, y_d)$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection into the first k coordinates and $Z = g\left(\frac{Y}{\|Y\|}\right) = \frac{1}{\|Y\|}(y_1, \ldots, y_k)$ and $L = \|Z\|^2$. It is clear that $\mathbb{E}L = \frac{k}{d}$. In fact, L is very concentrated around its mean*

- *If $\beta < 1$,*
$$\Pr\left[L \leq \beta\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1 - \beta + \log\beta)\right).$$
- *If $\beta > 1$,*
$$\Pr\left[L \geq \beta\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1 - \beta + \log\beta)\right).$$

Proof. [of Johnson-Lindenstrauss Lemma]

We will start by showing that, given a pair x_i, x_j a projection onto a random subspace of dimension k will satisfy (after appropriate scaling) property (48) with high probability. WLOG, we can assume that $u = x_i - x_j$ has unit norm. Understanding what is the norm of the projection of u on a random subspace of dimension k is the same as understanding the norm of the projection of a (uniformly)

random point on S^{d-1} the unit sphere in $\mathbb{R}^d$ on a specific $k-$dimensional subspace, let's say the one generated by the first k canonical basis vectors.

This means that we are interested in the distribution of the norm of the first k entries of a random vector drawn from the uniform distribution over S^{d-1} – this distribution is the same as taking a standard Gaussian vector in $\mathbb{R}^d$ and normalizing it to the unit sphere.

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection on a random $k-$dimensional subspace and let $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as $f = \frac{d}{k}g$. Then (by the above discussion), given a pair of distinct x_i and x_j, $\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2}$ has the same distribution as $\frac{d}{k}L$, as defined in Lemma 5.2. Using Lemma 5.2, we have, given a pair x_i, x_j,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1-\epsilon)+\log(1-\epsilon))\right),$$

since, for $\epsilon \geq 0$, $\log(1-\epsilon) \leq -\epsilon - \epsilon^2/2$ we have

$$\begin{aligned}\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\epsilon^2}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

On the other hand,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \geq (1+\epsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1+\epsilon)+\log(1+\epsilon))\right).$$

since, for $\epsilon \geq 0$, $\log(1+\epsilon) \leq \epsilon - \epsilon^2/2 + \epsilon^3/3$ we have

$$\begin{aligned}\text{Prob}\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\left(\epsilon^2-2\epsilon^3/3\right)}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

By union bound it follows that

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}.$$

Since there exist $\binom{n}{2}$ such pairs, again, a simple union bound gives

$$\Pr\left[\exists_{i,j} : \frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}\frac{n(n-1)}{2} = 1-\frac{1}{n}.$$

Therefore, choosing f as a properly scaled projection onto a random $k-$dimensional subspace is an $\epsilon-$ isometry on X (see (48)) with probability at least $\frac{1}{n}$. We can achieve any desirable constant probability of success by trying $\mathcal{O}(n)$ such random projections, meaning we can find an $\epsilon-$isometry in randomized polynomial time.

□

Note that by considering k slightly larger one can get a good projection on the first random attempt with very good confidence. In fact, it's trivial to adapt the proof above to obtain the following Lemma:

Lemma 5.3 *For any $0 < \epsilon < 1$, $\tau > 0$, and for any integer n, let k be such that*

$$k \geq (2+\tau)\frac{2}{\epsilon^2/2 - \epsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, take $f : \mathbb{R}^d \to \mathbb{R}^k$ to be a suitably scaled projection on a random subspace of dimension k, then f is an $\epsilon-$isometry for X (see (48)) with probability at least $1 - \frac{1}{n^\tau}$.

Lemma 5.3 is quite remarkable. Think about the situation where we are given a high-dimensional data set in a streaming fashion – meaning that we get each data point at a time, consecutively. To run a dimension-reduction technique like PCA or Diffusion maps we would need to wait until we received the last data point and then compute the dimension reduction map (both PCA and Diffusion Maps are, in some sense, data adaptive). Using Lemma 5.3 you can just choose a projection at random in the beginning of the process (all ones needs to know is an estimate of the log of the size of the data set) and just map each point using this projection matrix which can be done online – we don't need to see the next point to compute the projection of the current data point. Lemma 5.3 ensures that this (seemingly naïve) procedure will, with high probably, not distort the data by more than ϵ.

5.1.1 Optimality of the Johnson-Lindenstrauss Lemma

It is natural to ask whether the dependency on ϵ and n in Lemma 5.3 can be improved. Noga Alon [Alo03] showed that there are n points for which the smallest dimension k on which they can be embedded with a distortion as in Lemma 5.3, satisfies $k = \Omega\left(\frac{1}{\log(1/\epsilon)}\epsilon^{-2}\log n\right)$, this was recently improved by Larsen and Nelson [**?**], for linear maps, to $\Omega\left(\epsilon^{-2}\log n\right)$, closing the gap.[22]

5.1.2 Fast Johnson-Lindenstrauss

(Disclaimer: the purpose of this section is just to provide a bit of intuition, there is a lot of hand-waving!!)

Let's continue thinking about the high-dimensional streaming data. After we draw the random projection matrix, say M, for each data point x, we still have to compute Mx which, since M has $\mathcal{O}(\epsilon^{-2}\log(n)d)$ entries, has a computational cost of $\mathcal{O}(\epsilon^{-2}\log(n)d)$. In some applications this might be too expensive, can one do better? There is no hope of (significantly) reducing the number of rows (Recall Open Problem **??** and the lower bound by Alon [Alo03]). The only hope is to speed up the matrix-vector multiplication. If we were able to construct a sparse matrix M then we would definitely speed up the computation of Mx but sparse matrices tend to distort sparse vectors, and the data set may contain. Another option would be to exploit the Fast Fourier Transform and compute the Fourier Transform of x (which takes $\mathcal{O}(d\log d)$ time) and then multiply the Fourier Transform of x by a sparse matrix. However, this again may not work because x might have a sparse Fourier Transform. The solution comes from leveraging an uncertainty principle — it is impossible for both x and the FT of x to be sparse simultaneously. The idea is that if, before one takes the Fourier Transform of x, one flips (randomly) the signs of x, then the probably of obtaining a sparse vector is very small so a sparse matrix can be used for projection. In a nutshell the algorithm has M be a matrix of the form PHD,

[22] An earlier version of these notes marked closing the gap as an open problem, this has been corrected.

where D is a diagonal matrix that flips the signs of the vector randomly, H is a Fourier Transform (or Hadamard transform) and P a sparse matrix. This method was proposed and analysed in [AC09] and, roughly speaking, achieves a complexity of $\mathcal{O}(d \log d)$, instead of the classical $\mathcal{O}(\epsilon^{-2} \log(n) d)$.

There is a very interesting line of work proposing fast Johnson Lindenstrauss projections based on sparse matrices. In fact, this is, in some sense, the motivation for Open Problem 4.4. in [Ban15d]. We recommend these notes Jelani Nelson's notes for more on the topic [Nel].

5.2 Gordon's Theorem

In the last section we showed that, in order to approximately preserve the distances (up to $1 \pm \varepsilon$) between n points it suffices to randomly project them to $\Theta\left(\epsilon^{-2} \log n\right)$ dimensions. The key argument was that a random projection approximately preserves the norm of every point in a set S, in this case the set of differences between pairs of n points. What we showed is that, in order to approximately preserve the norm of every point in S it is enough to project to $\Theta\left(\epsilon^{-2} \log |S|\right)$ dimensions. The question this section is meant to answer is: can this improved if S has a special structure? Given a set S, what is the measure of complexity of S that explains how many dimensions one needs to take on the projection to still approximately preserve the norms of points in S. Was we will see below, this will be captured, via Gordon's Theorem, by the so called Gaussian Width of S.

Definition 5.4 (Gaussian Width) *Given a closed set $S \subset \mathbb{R}^d$, its* gaussian width $\omega(S)$ *is define as:*

$$\omega(S) = \mathbb{E} \max_{x \in S} \left[g_d^T x \right],$$

where $g_d \sim \mathcal{N}(0, I_{d \times d})$.

Similarly to what we did in the proof of Theorem 5.1 we will restrict our attention to sets S of unit norm vectors, meaning that $S \subset \mathbb{S}^{d-1}$.

Also, we will focus our attention not in random projections but in the similar model of random linear maps $G : \mathbb{R}^d \to \mathbb{R}^k$ that are given by matrices with i.i.d. gaussian entries. For this reason the following Proposition will be useful:

Proposition 5.5 *Let $g_k \sim \mathcal{N}(0, I_{k \times k})$, and define*

$$a_k := \mathbb{E} \|g_k\|.$$

Then $\sqrt{\frac{k}{k+1}} \sqrt{k} \leq a_k \leq \sqrt{k}$.

We are now ready to present Gordon's Theorem.

Theorem 5.6 (Gordon's Theorem [Gor88]) *Let $G \in \mathbb{R}^{k \times d}$ a random matrix with independent $\mathcal{N}(0,1)$ entries and $S \subset \mathbb{S}^{d-1}$ be a closed subset of the unit sphere in d dimensions. Then*

$$\mathbb{E} \max_{x \in S} \left\| \frac{1}{a_k} G x \right\| \leq 1 + \frac{\omega(S)}{a_k},$$

and

$$\mathbb{E}\min_{x\in S}\left\|\frac{1}{a_k}Gx\right\| \geq 1-\frac{\omega(S)}{a_k},$$

where $a_k = \mathbb{E}\|g_k\|$ *and* $\omega(S)$ *is the gaussian width of* S*. Recall that* $\sqrt{\frac{k}{k+1}}\sqrt{k} \leq a_k \leq \sqrt{k}$*.*

Before proving Gordon's Theorem we'll note some of it's direct implications. It suggest that $\frac{1}{a_k}G$ preserves the norm of the points in S up to $1 \pm \frac{\omega(S)}{a_k}$, indeed we can make this precise with Gaussian Concentration.

Note that the function $F(G) = \max_{x\in S}\left\|\frac{1}{a_k}Gx\right\|$ is 1-Lipschitz. Indeed

$$\begin{aligned}\left|\max_{x_1\in S}\|G_1x_1\| - \max_{x_2\in S}\|G_2x_2\|\right| &\leq \max_{x\in S}|\|G_1x\| - \|G_2x\|| \leq \max_{x\in S}\|(G_1-G_2)\,x\| \\ &= \|G_1-G_2\| \leq \|G_1-G_2\|_F\,.\end{aligned}$$

Similarly, one can show that $F(G) = \min_{x\in S}\left\|\frac{1}{a_k}Gx\right\|$ is 1-Lipschitz. Thus, one can use Gaussian Concentration to get:

$$\text{Prob}\left\{\max_{x\in S}\|Gx\| \geq a_k + \omega(S) + t\right\} \leq \exp\left(-\frac{t^2}{2}\right), \tag{49}$$

and

$$\text{Prob}\left\{\min_{x\in S}\|Gx\| \leq a_k - \omega(S) - t\right\} \leq \exp\left(-\frac{t^2}{2}\right). \tag{50}$$

This gives us the following Theorem.

Theorem 5.7 *Let* $G \in \mathbb{R}^{k\times d}$ *a random matrix with independent* $\mathcal{N}(0,1)$ *entries and* $S \subset \mathbb{S}^{d-1}$ *be a closed subset of the unit sphere in* d *dimensions. Then, for* $\varepsilon > \sqrt{\frac{\omega(S)^2}{a_k^2}}$*, with probability* $\geq 1 - 2\exp\left[-k\left(\varepsilon - \frac{\omega(S)}{a_k}\right)^2\right]$*:*

$$(1-\varepsilon)\|x\| \leq \left\|\frac{1}{a_k}Gx\right\| \leq (1+\varepsilon)\|x\|,$$

for all $x \in S$*.*

Recall that $k - \frac{k}{k+1} \leq a_k^2 \leq k$*.*

Proof. This is readily obtained by taking $\varepsilon = \frac{\omega(S)+t}{a_k}$, using (49), (50), and recalling that $a_k^2 \leq k$. □

Remark 5.8 *Note that a simple use of a union bound*[23] *shows that* $\omega(S) \lesssim \sqrt{2\log|S|}$*, which means that taking* k *to be of the order of* $\log|S|$ *suffices to ensure that* $\frac{1}{a_k}G$ *to have the Johnson Lindenstrauss property. This observation shows that Theorem 5.7 essentially directly implies Theorem 5.1 (although not exacly, since* $\frac{1}{a_k}G$ *is not a projection).*

[23] This follows from the fact that the maximum of n standard gaussian random variables is $\lesssim \sqrt{2\log|S|}$.

5.2.1 Gordon's Escape Through a Mesh Theorem

Theorem 5.7 suggests that, if $\omega(S) \leq a_k$, a uniformly chosen random subspace of $\mathbb{R}^n$ of dimension $(n-k)$ (which can be seen as the nullspace of G) avoids a set S with high probability. This is indeed the case and is known as Gordon's Escape Through a Mesh Theorem, it's Corollary 3.4. in Gordon's original paper [Gor88]. See also [Mix14b] for a description of the proof. We include the Theorem below for the sake of completeness.

Theorem 5.9 (Corollary 3.4. in [Gor88]) *Let $S \subset \mathbb{S}^{d-1}$ be a closed subset of the unit sphere in d dimensions. If $\omega(S) < a_k$, then for a $(n-k)$-dimensional subspace Λ drawn uniformly from the Grassmanian manifold we have*

$$\text{Prob}\left\{\Lambda \cap S \neq \emptyset\right\} \leq \frac{7}{2}\exp\left(-\frac{1}{18}\left(a_k - \omega(S)\right)^2\right),$$

where $\omega(S)$ is the gaussian width of S and $a_k = \mathbb{E}\|g_k\|$ where $g_k \sim \mathcal{N}(0, I_{k\times k})$.

5.2.2 Proof of Gordon's Theorem

In order to prove this Theorem we will use extensions of the Slepian's Comparison Lemma.

Slepian's Comparison Lemma, and the closely related Sudakov-Fernique inequality, are crucial tools to compare Gaussian Processes. A Gaussian process is a family of gaussian random variables indexed by some set T, $\{X_t\}_{t\in T}$ (if T is finite this is simply a gaussian vector). Given a gaussian process X_t, a particular quantity of interest is $\mathbb{E}\left[\max_{t\in T} X_t\right]$. Intuitively, if we have two Gaussian processes X_t and Y_t with mean zero $\mathbb{E}\left[X_t\right] = \mathbb{E}\left[Y_t\right] = 0$, for all $t \in T$, and the same variance, then the process that has the "least correlations" should have a larger maximum (think the maximum entry of vector with i.i.d. gaussian entries versus one always with the same gaussian entry). The following inequality makes this intuition precise and extends it to processes with different variances. [24]

Theorem 5.10 (Slepian/Sudakov-Fernique inequality) *Let $\{X_u\}_{u\in U}$ and $\{Y_u\}_{u\in U}$ be two (almost surely bounded) centered Gaussian processes indexed by the same (compact) set U. If, for every $u_1, u_2 \in U$:*

$$\mathbb{E}\left[X_{u_1} - X_{u_2}\right]^2 \leq \mathbb{E}\left[Y_{u_1} - Y_{u_2}\right]^2, \tag{51}$$

then

$$\mathbb{E}\left[\max_{u\in U} X_u\right] \leq \mathbb{E}\left[\max_{u\in U} Y_u\right].$$

The following extension is due to Gordon [Gor85, Gor88].

Theorem 5.11 *[Theorem A in [Gor88]] Let $\{X_{t,u}\}_{(t,u)\in T\times U}$ and $\{Y_{t,u}\}_{(t,u)\in T\times U}$ be two (almost surely bounded) centered Gaussian processes indexed by the same (compact) sets T and U. If, for every $t_1, t_2 \in T$ and $u_1, u_2 \in U$:*

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_1,u_2}\right]^2 \leq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_1,u_2}\right]^2, \tag{52}$$

[24] Although intuitive in some sense, this turns out to be a delicate statement about Gaussian random variables, as it does not hold in general for other distributions.

and, for $t_1 \neq t_2$,

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_2,u_2}\right]^2 \geq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_2,u_2}\right]^2, \tag{53}$$

then

$$\mathbb{E}\left[\min_{t\in T}\max_{u\in U} X_{t,u}\right] \leq \mathbb{E}\left[\min_{t\in T}\max_{u\in U} Y_{t,u}\right].$$

Note that Theorem 5.10 easily follows by setting $|T| = 1$.

We are now ready to prove Gordon's theorem.

Proof. [of Theorem 5.6]

Let $G \in \mathbb{R}^{k\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries. We define two gaussian processes: For $v \in S \subset \mathbb{S}^{d-1}$ and $u \in \mathbb{S}^{k-1}$ let $g \sim \mathcal{N}\left(0, I_{k\times k}\right)$ and $h \sim \mathcal{N}\left(0, I_{d\times d}\right)$ and define the following processes:

$$A_{u,v} = g^T u + h^T v,$$

and

$$B_{u,v} = u^T G v.$$

For all $v, v' \in S \subset \mathbb{S}^{d-1}$ and $u, u' \in \mathbb{S}^{k-1}$,

$$\begin{aligned}
\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 &= 4 - 2\left(u^T u' + v^T v'\right) - \sum_{ij}\left(v_i u_j - v_i' u_j'\right)^2 \\
&= 4 - 2\left(u^T u' + v^T v'\right) - \left[2 - 2\left(v^T v'\right)\left(u^T u'\right)\right] \\
&= 2 - 2\left(u^T u' + v^T v' - u^T u' v^T v'\right) \\
&= 2\left(1 - u^T u'\right)\left(1 - v^T v'\right).
\end{aligned}$$

This means that $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 \geq 0$ and $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 = 0$ if $v = v'$.

This means that we can use Theorem 5.11 with $X = A$ and $Y = B$, to get

$$\mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u} \leq \mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u}.$$

Noting that

$$\mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u} = \mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} u^T G v = \mathbb{E}\min_{v\in S}\|Gv\|,$$

and

$$\mathbb{E}\left[\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u}\right] = \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u + \mathbb{E}\min_{v\in S} h^T v = \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u - \mathbb{E}\max_{v\in S}(-h^T v) = a_k - \omega(S),$$

gives the second part of the Theorem.

On the other hand, since $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 \geq 0$ then we can similarly use Theorem 5.10 with $X = B$ and $Y = A$, to get

$$\mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u} \geq \mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u}.$$

Noting that

$$\mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u} = \mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{k-1}} u^T G v = \mathbb{E}\max_{v\in S}\|Gv\|,$$

and

$$\mathbb{E}\left[\max_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u}\right] = \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u + \mathbb{E}\max_{v\in S} h^T v = a_k + \omega(S),$$

concludes the proof of the Theorem.

□

5.3 Sparse vectors and Low-rank matrices

In this Section we illustrate the utility of Gordon's theorem by undertanding which projections are expected to keep the norm of sparse vectors and low-rank matrices.

5.3.1 Gaussian width of k-sparse vectors

Say we have a signal (or image) $x \in \mathbb{R}^N$ that we are interested in measuring with linear measurements $y_i = a_i^T x$, for $a_i \in \mathbb{R}^N$. In general, it is clear that we would need N measurements to find x. The idea behind *Compressed Sensing* [CRT06a, Don06] is that one may be able to significantly decrease the number of measurements needed if we know more about the structure of x, a prime example is when x is known to have few non-zero entries (being sparse). Sparse signals do arise in countless applications (for example, images are known to be sparse in the Wavelet basis; in fact this is the basis of the JPEG2000 compression method).

We'll revisit sparse recovery and Compressed Sensing next lecture but for now we'll see how Gordon's Theorem can suggest us how many linear measurements are needed in order to reconstruct a sparse vector. An efficient way of representing the measurements is to use a matrix

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_M^T & — \end{bmatrix},$$

and represent the linear measurements as

$$y = Ax.$$

In order to hope to be able to reconstruct x from y we need that A is injective on sparse vectors. Let us assume that x is s-sparse, meaning that x has at most s non-zero entries (often written as $\|x\|_0 \leq s$, where $\|\cdot\|_0$ is called the 0-norm and counts the number of non-zero entries in a vector[25]). It is also intuitive that, in order for reconstruction to be stable, one would like that not only A is injective in s-sparse vectors but actually almost an isometry, meaning that the ℓ_2 distance between Ax_1 and Ax_2 should be comparable to the distances between x_1 and x_2 if they are s-sparse. Since the difference between two s-sparse vectors is a $2s$-sparse vector, we can alternatively ask for A to keep the norm of $2s$ sparse vectors. Gordon's Theorem above suggests that we can take $A \in \mathbb{R}^{M\times N}$ to have

[25] It is important to note that $\|\cdot\|_0$ is not actually a norm

i.i.d. gaussian entries and to take $M \approx \omega(\mathcal{S}_{2s})$, where $\mathcal{S}_k = \left\{x : x \in \mathbb{S}^{N-1}, \|x\|_0 \leq k\right\}$ is the set of $2s$ sparse vectors, and $\omega(\mathcal{S}_{2s})$ the gaussian width of $\mathcal{S}_{2s}$.

Proposition 5.12 *If $s \leq N$, the Gaussian Width $\omega(\mathcal{S}_s)$ of $\mathcal{S}_s$, the set of unit-norm vectors that are at most s sparse, satisfies*

$$\omega(\mathcal{S}_s) \lesssim s \log\left(\frac{N}{s}\right).$$

Proof.

$$\omega(\mathcal{S}_s) = \max_{v \in SS^{N-1},\, \|v\|_0 \leq s} g^T v, \log\left(\frac{N}{s}\right),$$

where $g \sim \mathcal{N}(0, I_{N\times N})$. We have

$$\omega(\mathcal{S}_s) = \max_{\Gamma \subset [N],\, |\Gamma| = s} \|g_\Gamma\|,$$

where g_Γ is the restriction of g to the set of indices Γ.

Given a set Γ, Theorem 4.12 gives

$$\text{Prob}\left\{\|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + t\right\} \leq \exp(-t).$$

Union bounding over all $\Gamma \subset [N]$, $|\Gamma| = s$ gives

$$\text{Prob}\left\{\max_{\Gamma \subset [N],\, |\Gamma| = s} \|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + t\right\} \leq \binom{N}{s} \exp(-t)$$

Taking u such that $t = su$, gives

$$\text{Prob}\left\{\max_{\Gamma \subset [N],\, |\Gamma| = s} \|g_\Gamma\|^2 \geq s\left(1 + 2\sqrt{u} + u\right)\right\} \leq \exp\left[-su + s\log\left(\frac{N}{s}\right)\right]. \tag{54}$$

Taking $u > \log\left(\frac{N}{s}\right)$ it can be readily seen that the typical size of $\max_{\Gamma \subset [N],\, |\Gamma| = s} \|g_\Gamma\|^2$ is $\lesssim s\log\left(\frac{N}{s}\right)$. The proof can be finished by integrating (54) in order to get a bound of the expectation of $\sqrt{\max_{\Gamma \subset [N],\, |\Gamma| = s} \|g_\Gamma\|^2}$.

□

This suggests that $\approx 2s\log\left(\frac{N}{2s}\right)$ measurements suffice to identify a $2s$-sparse vector. As we'll see, not only such a number of measurements suffices to identify a sparse vector but also for certain efficient algorithms to do so.

5.3.2 The Restricted Isometry Property and a couple of open problems

Matrices perserving the norm of sparse vectors do play a central role in sparse recovery, they are said to satisfy the Restricted Isometry Property. More precisely:

Definition 5.13 (The Restricted Isometry Property) *An $M \times N$ matrix A (with either real or complex valued entries) is said to satisfy the (s, δ)-Restricted Isometry Property (RIP),*

$$(1 - \delta)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta)\|x\|^2,$$

for all s-sparse x.

Using Proposition 5.12 and Theorem 5.7 one can readily show that matrices with Gaussian entries satisfy the restricted isometry property with $M \approx s \log\left(\frac{N}{s}\right)$.

Theorem 5.14 *Let A be an $M \times N$ matrix with i.i.d. standard gaussian entries, there exists a constant C such that, if*

$$M \geq Cs \log\left(\frac{N}{s}\right),$$

then $\frac{1}{a_M} A$ satisfies the $\left(s, \frac{1}{3}\right)$-RIP, with high probability.

Theorem 5.14 suggests that RIP matrices are abundant for $s \approx \frac{M}{\log(N)}$, however it appears to be very difficult to deterministically construct matrices that are RIP for $s \gg \sqrt{M}$, known as the square bottleneck [Tao07, BFMW13, BFMM14, BMM14, B$^+$11, Mix14a]. The only known unconditional construction that is able to break this bottleneck is due to Bourgain et al. [B$^+$11] that achieves $s \approx M^{\frac{1}{2}+\varepsilon}$ for a small, but positive, ε. There is a conditional construction, based on the Paley Equiangular Tight Frame, that will be briefly described in the next Lecture [BFMW13, BMM14].

Open Problem 5.1 *Construct deterministic matrices $A \in \mathbb{C}^{M \times N}$ (or $A \in \mathbb{C}^{M \times N}$) satisfying $(s, \frac{1}{3})$-RIP for $s \gtrsim \frac{M^{0.6}}{\text{polylog}(N}$.*

Open Problem 5.2 *Theorem 5.14 guarantees that if we take A to have i.i.d. Gaussian entries then it should be RIP for $s \approx \frac{M}{\log(N)}$. If we were able to, given A, certify that it indeed is RIP for some s then one could have a randomized algorithm to build RIP matrices (but that is guaranteed to eventually find one). This motives the following question*

1. *Let $N = 2M$, for which s is there a polynomial time algorithm that is guaranteed to, with high probability, certify that a gaussian matrix A is $\left(s, \frac{1}{3}\right)$-RIP?*
2. *In particular, a $\left(s, \frac{1}{3}\right)$-RIP matrix has to not have s sparse vectors in its nullspace. This motivates a second question: Let $N = 2M$, for which s is there a polynomial time algorithm that is guaranteed to, with high probability, certify that a gaussian matrix A does not have s-sparse vectors in its nullspace?*

The second question is tightly connected to the question of sparsest vector on a subspace (for which $s \approx \sqrt{M}$ is the best known answer), we refer the reader to [SWW12, QSW14, BKS13b] for more on this problem and recent advances. Note that checking whether a matrix has RIP or not is, in general, NP-hard [BDMS13, TP13].

5.3.3 Gaussian width of rank-r matrices

Another structured set of interest is the set of low rank matrices. Low-rank matrices appear in countless applications, a prime example being the Netflix Prize. In that particular example the matrix in question is a matrix indexed by users of the Netflix service and movies. Given a user and a movie, the corresponding entry of the matrix should correspond to the score that user would attribute to that movie. This matrix is believed to be low-rank. The goal is then to estimate the score for user and

movie pairs that have not been rated yet from the ones that have, by exploiting the low-rank matrix structure. This is known as low-rank matrix completion [CT10, CR09, Rec11].

In this short section, we will not address the problem of matrix completion but rather make a comment about the problem of low-rank matrix sensing, where instead of observing some of the entries of the matrix $X \in \mathbb{R}^{n_1 \times n_2}$ one has access to linear measuremetns of it, of the form $y_i = \mathrm{Tr}(A_i^T X)$.

In order to understand the number of measurements needed for the measurement procedure to be a nearly isometry for rank r matrices, we can estimate the Gaussian Width of the set of matrices $X \in\in \mathbb{R}^{n_1 \times n_2}$ whose rank is smaller or equal to $2r$ (and use Gordon's Theorem).

Proposition 5.15

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r\right\}\right) \lesssim \sqrt{r(d_1 + d_2)}.$$

Proof.

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r\right\}\right) = \mathbb{E} \max_{\substack{\|X\|_F = 1 \\ \mathrm{rank}(X) \leq r}} \mathrm{Tr}(GX).$$

Let $X = U\Sigma V^T$ be the SVD decomposition of X, then

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r\right\}\right) = \mathbb{E} \max_{\substack{U^T U = V^T V = I_{r \times r} \\ \Sigma \in \mathbb{R}^{r \times r} \text{ diagonal } \|\Sigma\|_F = 1}} \mathrm{Tr}(\Sigma\left(V^T G U\right)).$$

This implies that

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r\right\}\right) \leq \left(\mathrm{Tr}\,\Sigma\right)\left(\mathbb{E}\|G\|\right) \lesssim \sqrt{r}\left(\sqrt{n_1} + \sqrt{n_1}\right),$$

where the last inequality follows from bounds on the largest eigenvalue of a Wishart matrix, such as the ones used on Lecture 1. □

18.S096 Topics in Mathematics of Data Science
Fall 2015