Finding a sparse vector in a subspace

The *sparsest vector problem* is the following:

- **Input:** A subspace $V \subseteq \mathbb{R}^n$ of dimension $k+1$ (given in the form of a basis)
- **Guarantee:** There exists some $v_0 \in V$ with at most ϵn nonzero coordinates.
- **Goal:** Find some vector $v' \in V$ that is nearly sparse, in the sense that $\|v' - \Pi v'\| \ll \|v'\|$ where $\Pi v'$ is the projection of v' to its ϵn largest coordinates.

This problem makes sense in both the *worst case* and *average case* settings but our interest in this lecture will be mostly in the latter, where the subspace V is obtained by taking k random vectors $v_1, \ldots, v_k$ and letting $V = Span\{v_0, \ldots, v_k\}$in The input is an arbitrary basis for V (or, if you want, $k+1$ samples from the basis-independent standard Gaussian distribution over vectors $v \in V$).

The problem itself is somewhat natural, and can be thought of as an average-case real (as opposed to finite field) version of the "shortest codeword" or "lattice shortest vector" problem. This also turns out to be related (at least in terms of techniques) to problems in unsupervised learning such as dictionary learning / sparse coding.

There is a related problem, often called "compressed sensing" or "sparse recovery" in which we are given an *affine* subspace A of the form $v_0 + V$, where v_0 is again sparse and V is an (essentially) random linear subspace, and the goal is again to recover v_0. Note that typically this problem is described somewhat differently: we have an $m \times n$ matrix A, often chosen at random, and we get the value $y = Av_0$. This determines the $k = n - m$ dimensional affine subspace $v_0 + \ker(A)$, and we need to recover v_0.

One difference between the problems is parameters (we will think of $k \ll n$, while in sparse recovery typically $k \sim n - o(n)$), but another more fundamental difference is that a linear subspace always has the all-zeroes vector in it, and hence, in contrast to the affine case, v_0 is *not* the sparsest vector in the subspace (only the sparsest nonzero one).

This complicates matters, as the algorithm of choice for sparse recovery is L1 minimization: find $v \in A$ that minimizes $\|v\|_1 =$

$\sum_{i=1}^{n} |v_i|$. This can be done by solving the linear program:

$$\begin{aligned} & \min \sum_{i=1}^{n} x_i \\ \text{subject to} \quad & x_i \geq v_i \\ & x_i \geq -v_i \\ & v \in A \end{aligned} \tag{1}$$

But of course if A were a linear subspace but not affine, then this would return the all-zero vector. (Though see below on variants that do make sense for the planted vector problem.)

Formal description of average case problem

We assume that $v_1, \ldots, v_k \in \mathbb{R}^n$ are chosen randomly as standard Gaussian vectors (i.e. with i.i.d. entries drawn from $N(0,1)$), and v_0 is some arbitrary unit vector with at most εn nonzero coordinates. We are given an arbitrary basis B for $Span\{v_0, v_1, \ldots, v_k\}$. The goal is to recover v_0.

For this lecture, this means recovering a unit vector v such that $\langle v, v_0 \rangle^2 \geq 0.99$ (though see Barak et al. [2014] for recovery with arbitrary accuracy). For simplicity let's also assume that v_0 is orthogonal to $v_1, \ldots, v_k$. (This is not really needed but helps simplify some minor calculations.)

Ratios of Norms

Rather than trying to directly trying to find a sparse vector, we will define a smoother *proxy* for sparsity, that is some polynomial $P(\cdot)$ so that $P(v)$ is larger for sparse vectors than for small ones. Then we will look for a vector v in the subspace that maximizes $P(v)$ (subject to some normalization) and hope that (a) we can efficiently do this and (b) that the answer is v_0. This makes the problem more amenable for the SOS algorithm and also makes for a more robust notion, allowing for some noise in v_0 (and allows us to not worry so much about issues of numerical accuracy).

So, we want some function that will favor vectors that are "spikier" as opposed to "smoother". We use the observation that taking high powers amplifies "spikes". Specifically, we note that if $q > p$ a sparse/spiky vector v would have a larger ratio of $\|v\|_q / \|v\|_p$ than

[1] For $p > 0$, the *p-norm* of a vector v, denoted as $\|v\|_p$ is defined as $(\sum_i |v_i|^p)^{1/p}$. By taking limits one can also define $\|v\|_\infty = \max_i |v_i|$ and $\|v\|_0 = |\{i \mid v_i \neq 0\}|$.

a dense/smooth one.[1] Indeed, compare the all 1's vector $\vec{1}$ with the vector 1_S for a set S of size εn. $\|\vec{1}\|_q / \|\vec{1}\|_p = n^{1/q-1/p}$ while $\|1_S\|_q / \|1_S\|_p = (\varepsilon n)^{1/q-1/p}$ which means that if $q > p$, the latter ratio is larger than the former by some power of $1/\varepsilon$. Moreover, an application of Hölder's inequality reveals that if v is εn-sparse then its q vs p norm ratio can only be higher than this:

1. Lemma. *If $v \in \mathbb{R}^n$ has at most ϵn nonzero coordinates, then*

$$(\mathbb{E}\, i v(i)^q)^{1/q} \geq \varepsilon^{1/q-1/p} (\mathbb{E}\, i v(i)^p)^{1/p}. \tag{2}$$

Proof. Let $1_{|v|>0}$ be the vector which is 1 if $|v(i)| > 0$ and 0 otherwise. Let $w \in \mathbb{R}^n$ be given by $w = 1_{|v|>0}/n^{1-q/p}$. Then by Hölder's inequality,

$$\begin{aligned}(\mathbb{E}\, i v(i)^p) &= \sum_i w(i) \frac{v(i)^p}{n^{q/p}} \\ &\leq (\sum_i w(i)^{1/(1-p/q)})^{1-p/q} (\sum_i v(i)^q / n)^{p/q} \\ &= \varepsilon^{1-p/q} (\mathbb{E}\, i v(i)^q)^{p/q}.\end{aligned} \tag{3}$$

Rearranging gives the result. □

How good a proxy for sparsity is this? We know that vectors which are actually sparse "look sparse" in the ratio-of-norms sense, but what about the other way around: could the ratio of norms measure be "fooled" by vectors which are not actually sparse? The answer is yes. For example, if $q = \infty$ and $p = 1$, the vector which has a 1 in one coordinate and ε in the other coordinates looks like an ϵ-sparse (or more accurately $\varepsilon - 1/n$-sparse) vector as far as the ∞ versus 1 norm ratio is concerned, but in the strict ℓ_0-sense is actually maximally non-sparse.

However, as the gap between p and q shrinks, a random subspace becomes less and less likely to contain these kind of "cheating vectors" that are not sparse but look sparse when comparing ℓ_q versus ℓ_p norms. Alternatively phrased, the closer we can take p and q, the higher dimension random subspace we can tolerate before the subspace becomes likely to contain a vector which confuses the ℓ_q versus ℓ_p sparsity proxy.

Unfortunately, there are very few values $q > p$ for which we know how to compute $\max_{v \in V} \|v\|_q / \|v\|_p$.[2] Demanet and Hand **?** and Spielman, Wang, and Wright **?** use the ℓ_∞ versus ℓ_1 proxy for sparsity to attack this problem). This can be computed efficiently (see exercise below) but if $k \gg 1$, this will not detect a vector v that is 0.01-sparse.

[2] Some examples include $q = \infty$ and $p \in \{1, 2\}$ (**exercise**), see also Bhaskara and Vijayaraghavan [2011] .

2. Exercise. Give an efficient algorithm to compute $\min \|v\|_1$ over all $v \in V$ with $\max |v_i| \geq 1$.[3]

3. Exercise. Give an efficient algorithm to compute $\max_{\|v\|_2=1} \|v\|_infty$.[4]

4. Exercise. Prove that for every subspace V of dimension k, there exists a vector $v \in V$ with $\max_i v_i = 1$ and $\sum |v_i| \leq \sqrt{k}/(10n)$

Some works have suggested to use the ℓ_2 vs ℓ_1 proxy. Which actually works pretty well in the sense that if V is a random subspace of dimension at most ηn, then there is no vector $v \in V$ whose ℓ_2 vs ℓ_1 ratio pretends to be a δ-sparse vector where δ is some function of η.

5. Exercise. Prove that for every $\eta < 1$ there exists some $\delta = \delta(\eta)$ such that if $v_1, \ldots, v_{\eta n}$ are random Standard Gaussian vectors (each coordinate is distributed according to $N(0,1)$) then with probability at least 0.9 for every $x \in \mathbb{R}^{\varepsilon n}$ with $\|x\|_2^2 = 1$

$$\sum_{i=1}^{\varepsilon n} |\langle v_i, x\rangle| \geq \delta n \tag{4}$$

6. Exercise. Using the above, show that for every $\eta < 1$, there is some $\delta = \delta(\eta)$ such that a random subspace (in our model above) does not contain a δ-sparse vector.

However, the ℓ_2 vs ℓ_1 problem has one caveat - we don't know how to compute it, even for a random subspace. In fact, this problem seems quite related to the question of certifying the *restricted isometry property* of a matrix— this is the goal of certifying the a random $m \times n$ matrix A (for $n > m$) satisfies that $\|Ax\|_2 \in (C, 1/C)\|x\|_2$ for every *sparse* vector x. In particular this would be false if there was a sparse vector in the *Kernel* of A, which is a subspace of $\mathbb{R}^n$ of dimension $m - n$. Known methods to certify this property require that the sparse vector x has at most $\sqrt{m}$ nonzero coordinates. See also this blog post of Tao and Koiran and Zouzias [2014] .

The 2 to 4 norm problem

In the following, we will use ℓ_4 versus ℓ_2 as our proxy for sparsity. It might seem like a strange choice since a priori it appears to yield the "worst of both worlds". On one hand, though it is better than the ℓ_∞ vs ℓ_1 proxy, the ℓ_4/ℓ_2 ratio is a worse proxy than the ℓ_2 vs ℓ_1 ratio, and to detect $1/100$-sparse vectors we will need to require the dimension k of the subspace to be at most $\varepsilon\sqrt{n}$ for some $\varepsilon > 0$ (which is much better than $k = O(1)$ needed in the ℓ_∞/ℓ_1 case but

[3] **Hint:** First show that for every i, minimizing $\|v\|_1$ over $v \in V$ where $v_i = 1$ can be done efficiently via a linear program.

[4] **Hint:** Use the fact that this is the same as maximizing over all i $\langle e_i, Ax\rangle$ where A is an $m \times n$ matrix whose columns are an orthonormal basis for V.

$k = \Omega(n)$ achieved in the ℓ_2/ℓ_1 case). On the other hand, we don't know how to compute this ratio either. In fact, Barak et al. [2012] showed (via connections with the quantum separability problem) that computing this ratio cannot be done in $n^{O(\log n)}$ time unless SAT has a subexponential time algorithm, and that even achieving weaker approximations would break the Small-Set Expansion (and hence probably also the Unique Games) conjecture. Nevertheless, we will show that we can in fact compute this ratio in the random case, using the degree 4 SOS system.

7. Exercise. Show that the 2 to 4 ratio cannot detect $1/100$-sparse vectors if the subspace has dimension much larger than $\sqrt{n}$. That is, prove that if $V \subseteq R^n$ has dimension $k > \sqrt{n}$ then there is a vector $v \in V$ such that $\mathbb{E}\, v_i^4 \geq \frac{k^2}{10n} \left(\mathbb{E}\, v_i^2\right)^2$.

Sparsest vector via sos

Maximizing the 2 to 4 norm over a subspace $V \subseteq \mathbb{R}^m$ of dimension n can be phrased as the polynomial optimization problem $\max_{\|x\|_2=1} \|Bx\|_4^4$ where B is the $m \times n$ generating matrix for the subspace V (i.e. $Im(B) = V$ and $\|Bx\|_2 = \|x\|_2$ for all x). We run the degree 4 sos algorithm to obtain a pseudo-distribution μ over the sphere. The rounding algorithm will simply be to use the *quadratic sampling lemma* to sample a random w from a Gaussian distribution whose second moments match those of μ. Thus, analyzing this algorithm boils down to proving the following:

8. Theorem (Sparse vector recovery). *If the subspace $V = \mathrm{Span}\{v_1, \ldots, v_k\}$ is chosen at random and v_0 is a unit vector orthogonal to $v_1, \ldots, v_k$ and has at most $0.00001k^2$ nonzero coordinates, then for every distribution μ over unit vectors in V where $\tilde{\mathbb{E}}_{\mu(w)} \|w\|_4^4 = \|v_0\|_4^4$ $\tilde{\mathbb{E}}_{\mu} \|Pw\|_2^2 \leq 0.01$ where P is the projector to $\mathrm{Span}\{v_1, \ldots, v_k\}$.*

This result means that if $w \in V$ is a vector such that both $\|w\|^2$ and $\|Pw\|_2^2$ are close to their expectations (which are 1 and at most 0.01 respectively) then, writing $w = \langle w, v_0 \rangle v_0 + w'$ where w' is in the span of $\{v_1, \ldots, v_k\}$, we see that $\|w'\|^2 \leq 0.01$ and hence $\langle w, v_0 \rangle^2 \geq 0.99$. Somewhat cumbersome but not too hard calculations spelled out below will show that we can get sufficiently close concentration (essentially since we can repeat the process and output the sparsest vector w we can find).

The algorithm described above only looks at the first two moments of the pseudo distribution. So, why did we need it to be a degree 4 (as opposed to degree 2) pseudo distribution? This is only for the proof, though note that the ℓ_4/ℓ_2 SOS program doesn't even make for degree < 4 pseudo-distributions.

Proof of Theorem 8

The SoS algorithm gives us a pseudodistribution that "pretends" to be supported on unit vectors $v \in Span\{v_0, \ldots, v_k\}$ such that $\|v\|_4^4 = C^4/n$. We first prove the main lemma for actual distributions and then demonstrate an instance of "Marley's Hypothesis": if you proved it for real distributions and didn't use anything too fancy, then every little thing gonna be all right (when you try to prove it for pseudodistributions).

The main result we will take at the moment as a given is the following:

9. Lemma (Random subspaces don't contain 2 to 4-sparse vectors). *If $k \ll \sqrt{n}$, with high probability*

$$\|Pv\|_4^4 \leq 10\|Pv\|_2^4/n \tag{5}$$

for every v.

We will show that Lemma 9 implies our Main Lemma for actual distributions. Namely, we show the following:

10. Lemma (2 to 4 sparsity implies correlation). *If P satisfies Eq. (5) then for every unit vector $w \in V$ with $\|w\|_4 \geq \|v_0\|_4/100 = C/100n^{1/4}$, the square correlation of w with v_0 satisfies $\langle w, v_0\rangle^2 \geq 1 - O(1/C)$.*

Proof. Let $w \in V$ be a unit vector. We can write $w = \alpha v_0 + Pw$. Hence, using the triangle inequality for the ℓ_4-norm,

$$\|w\|_4 \leq \alpha\|v_0\|_4 + \|Pw\|_4 \tag{6}$$

which can be rearranged to

$$\alpha \geq 1 - \frac{\|Pw\|_4}{\|v_0\|_4} \tag{7}$$

But since $\|v_0\|_4 = C/n^{1/4}$, and Lemma [lem:random:actual] $\|Pw\|_4 \leq 2/n^{1/4}$, the RHS is at least $1 - 2/C$. □

Lemma 10 concludes the proof of the main lemma in the actual distribution case since $\|w\|_2^2 = \langle w, v_0\rangle^2 + \|Pw\|_2^2$.

Pseudo-distribution version and proofs

We now state the pseudo-distribution versions of our lemmas and prove them:

11. Lemma (Random subspaces don't pseudo contain 2 to 4 sparse vectors). *With high probability*

$$\|Pv\|_4^4 \preceq 10\|Pv\|_2^4/n \tag{8}$$

where we now think of $\|Pv\|_4^4$ and $\|Pv\|_2^4$ as polynomials in indeterminates v and with coefficients determined by P, and $\preceq$ denotes that the polynomial $10\|Pv\|_2^4 - \|Pv\|_4^4$ is a sum of squares.

12. Lemma (Pseudo-sparsity implies correlation). *If P satisfies Eq. (8) then for every degree 4 pseudo-distribution over the unit sphere satisfying $\|x\|_4^4 = \|v_0\|_4^4 = C^4/n$ it holds that $\tilde{\mathbb{E}}\,\langle x, v_0\rangle^2 \geq 1 - O(1/C)$.*

Now we test "Marley's Hypothesis" by lifting the proof of Lemma 10 to the pseudo-distribution case and proving Lemma 12. We need to be able to mimic all the steps we used when everything is wrapped in pseudoexpectations. The main interesting step in the proof of Lemma 10 was our use of the *triangle inequality* that $\|x+y\|_4 \leq \|x\|_4 + \|y\|_4$.

13. Lemma (Triangle Inequality for Pseudodistributions). *Let μ be a degree-4 pseudodistribution over $\mathbb{R}^{2n}$. Then*

$$\tilde{\mathbb{E}}_{\mu(x,y)} \|x+y\|_4^{4^{1/4}} \leq \tilde{\mathbb{E}}\,\|x\|_4^{4^{1/4}} + \tilde{\mathbb{E}}\,\|y\|_4^{4^{1/4}}. \tag{9}$$

14. Exercise. Prove Lemma 13

We note that the following easier bound would be fine for us:

15. Exercise. If a pseudo-distribution μ over $(x,y) \in \mathbb{R}^{2n}$ satisfies $\tilde{\mathbb{E}}\|x\|_4^4 \geq \tilde{\mathbb{E}}\|y\|_4^4$ then

$$\tilde{\mathbb{E}}\,\|x+y\|_4^4 \leq \tilde{\mathbb{E}}\,\|x\|_4^4 + 15\left(\tilde{\mathbb{E}}\,\|x\|_4^{4^{1/4}}\right)^{3/4}\left(\tilde{\mathbb{E}}\,\|y\|_4^4\right)^{1/4}. \tag{10}$$

Proof of Lemma 12 from Lemma 11: The proof is almost identical to the proof of Lemma [lem:cor:actual]. Let P satisfy

$$\|Px\|_4^4 \preceq \frac{10\|Px\|_2^4}{n} \tag{11}$$

where we interpret both sides as polynomials in x. Let $\{x\}$ be a degree-4 pseudodistribution satisfying $\{\|x\|_2^2 = 1, \|x\|_4^4 = \|v_0\|_4^4 =$

$C^4/n\}$. Using the pseudodistribution triangle inequality,

$$\tilde{\mathbb{E}}\,\|x\|_4^{4^{1/4}} \le \tilde{\mathbb{E}}\,\|\langle x,v_0\rangle v_0\|_4^{4^{1/4}} + \tilde{\mathbb{E}}\,\|Px\|_4^4 = \frac{C}{n^{1/4}}\,\tilde{\mathbb{E}}\,\langle x,v_0\rangle^{4^{1/4}} + \tilde{\mathbb{E}}\,\|Px\|_4^{4^{1/4}}. \tag{12}$$

Rearranging and using our assumptions on $\{x\}$,

$$\tilde{\mathbb{E}}\,\langle x,v_0\rangle^{4^{1/4}} \ge \frac{n^{1/4}}{C}(\tilde{\mathbb{E}}\,\|x\|_4^{4^{1/4}} - \tilde{\mathbb{E}}\,\|Px\|_4^{4^{1/4}}) = 1 - \frac{n^{1/4}}{C}\,\tilde{\mathbb{E}}\,\|Px\|_4^{4^{1/4}}. \tag{13}$$

Now we use our assumption on P to get

$$\tilde{\mathbb{E}}\,\|Px\|_4^{4^{1/4}} \le 2\frac{\tilde{\mathbb{E}}\,\|Px\|_2^{4^{1/4}}}{n^{1/4}}. \tag{14}$$

Moreover, note that $\|Px\|_2^4 \preceq \|x\|_2^4$, since both are homogeneous degree-4 polynomials all of whose monomials are squares and the coefficient of every monomial on the left-hand side is smaller than the corresponding coefficient on the right. This gives

$$\tilde{\mathbb{E}}\,\|Px\|_2^4 \le \tilde{\mathbb{E}}\,\|x\|_2^4. \tag{15}$$

Putting it together, we get

$$\tilde{\mathbb{E}}\,\langle x,v_0\rangle^{4^{1/4}} \ge 1 - \frac{2}{C}\,\tilde{\mathbb{E}}\,\|x\|_2^{4^{1/4}}. \tag{16}$$

Since $\{x\}$ satisfies $\tilde{\mathbb{E}}\,\|x\|_2^2 = 1$, we have

$$\tilde{\mathbb{E}}\,\|x\|_2^2\left(\|x\|_2^2 - 1\right) = 0 \tag{17}$$

and therefore $\tilde{\mathbb{E}}\,\|x\|_2^4 = 1$. Plugging this in to the above,

$$\tilde{\mathbb{E}}\,\langle x,v_0\rangle^{4^{1/4}} \ge 1 - \frac{2}{C}. \tag{18}$$

The last step is to relate $\tilde{\mathbb{E}}\,\langle x,v_0\rangle^4$ and $\tilde{\mathbb{E}}\,\langle x,v_0\rangle^2$. Again using that $\{x\}$ satisfies $\tilde{\mathbb{E}}\,\|x\|_2^2 = 1$, we have

$$\tilde{\mathbb{E}}\,\langle x,v_0\rangle^2\|x\|_2^2 = \tilde{\mathbb{E}}\,\langle x,v_0\rangle^2. \tag{19}$$

Moreover, since $\langle x,v_0\rangle^2 \preceq \|x\|_2^2$ we must have $\langle x,v_0\rangle^4 \preceq \langle x,v_0\rangle^2\|x\|_2^2$ (the difference of the two sides in the former is a sum of squares; multiplying that SoS polynomial by the square polynomial $\|x,v_0\|^2$ yields another SoS polynomial which is the difference between the two sides in the latter case).

All together, we get

$$\tilde{\mathbb{E}}\,\langle x,v_0\rangle^2 \ge \tilde{\mathbb{E}}\,\langle x,v_0\rangle^4 \ge \left(1 - \frac{2}{C}\,\tilde{\mathbb{E}}\,\|x\|_2^{4^{1/4}}\right)^4 \ge 1 - \frac{8}{C} \tag{20}$$

and we are done.

Proof of Lemma 11

True to form, we would like to start by proving Lemma 9 and then lift the proof to the SoS setting. Lets start with a heuristic argument on why would Lemma 9 be true. Think of the case that we fix a unit vector $x \in \mathbb{R}^k$ and pick $v_1, \ldots, v_k$ as random Gaussian vectors of unit norm in $\mathbb{R}^n$, i.e., each entry is distributed as $N(0, 1/\sqrt{n})$. Then, the vector $w = \sum x_i v_i$ would have each coordinate be a Gaussian random variable distributed as $N(0, 1/\sqrt{n})$ (since $\sum x_i^2 = 1$). Now the probability $\|w\|_4^4 \geq C^4/n$ is the probability that $\sum_{i=1}^n g_i^4 \geq nC^4$ where the g_i's are independent standard Gaussians.

Typically a sum of independent random variables can be large for two different reasons. Either every one of those random variables is moderately large, or one of them is very large. In this case, the probability that every one of the g_i's would be of magnitude at least C is $\exp(-C^2)$ and so the probability that all of them satisfy this would be $\exp(-\Omega(n))$. This is much smaller than the probability that a single one of the g_i's has magnitude at least $Cn^{1/4}$ which is $\exp(-O(\sqrt{n}))$ and indeed one can show that the latter event is the one dominating this probability. Thus, if $C^2\sqrt{n} \gg k$, we can do a union bound over a sufficiently fine net of $\mathbb{R}^k$ and rule this out.

This argument can be turned into a proof, but note that we have used a concentration and union bound type of argument, i.e. the dreaded *probabilistic method*, and hence cannot appeal to Marley's Corollary for help. So, we will want to try to present a different argument, that still uses concentration but somehow will work out fine.

Intuition and Heuristic Argument

A formulation that will work just as well for the proof of the main theorem is: given an orthonormal basis matrix B for $\mathrm{Span}\{v_1, \ldots, v_k\}$,

$$\|Bv\|_4^4 \leq 10\|v\|_2^4/n \tag{21}$$

Now, the matrix B whose columns are $v_1/\sqrt{n}, \ldots, v_k/\sqrt{n}$ is almost such a matrix (since these vectors are random, they are nearly orthogonal), and so let's just assume it is the basis matrix. So, we need to show that if B has i.i.d. $N(0, 1/\sqrt{n})$ coordinates and $n \gg k^2$ then with high probability Eq. (21) is satisfied.

Let $w_1, \ldots, w_n$ be the rows of B. Then,

$$n\|Bv\|_4^4 = \sum_{i=1}^{n} \langle w_i, v\rangle^4 = \tfrac{1}{n} \sum_{i=1}^{n} n^2 \langle w_i, v\rangle^4 \tag{22}$$

That means that we can think of the polynomial $Q(v) = \|Bv\|_4^4 n$ as the average of n random polynomials each chosen as $\langle g, v\rangle^4$, where $g = \sqrt{n}w$ has i.i.d $N(0,1)$ entries. Since in expectation $\langle g, v\rangle^4 \leq 5\|v\|_2^4$ (**exercise**), we can see that if n is sufficiently large then $Q(v)$ will with high probability be very close to its expectation and so have $Q(v) \leq 10\|v\|_2^4$. It turns out that "sufficiently large" in this case means as long as $n \gg k^2$. Moreover, we will be able to show that in this case, $Q(v) = 10\|v\|_2^4 - s(v)$ where s is a *sum of squares* polynomial of degree four.

We now give some high level arguments on how to make this into a proper proof. We first recall the following exercise:

16. Exercise. Let P, Q be two homogenous n-variate degree 4 polynomials, and write $P \preceq Q$ if $Q - P$ is a sum of squares. Prove that $P \preceq Q$ if and only if there exist matrices M_P, M_Q such that for every $x \in \mathbb{R}^n$, $P(x) = \langle M_P, x^{\otimes 4}\rangle$ and $Q(x_= \langle M_Q, x^{\otimes 4}\rangle$ such that $M_P \preceq M_Q$ in the spectral sense. (i.e., where we say that a matrix A satisfies $A \preceq B$ if $w^\top A w \leq w^\top B w$ for all w.)

17. Exercise. Prove that a degree four polynomial P satisfies $P \preceq \lambda\|x\|_2^4$ if and only if there exists such a matrix M_P with $\|M_P\| \leq \lambda$ where $\|M_P\|$ denotes the spectral norm.[5]

This connection suggests using the *Matrix Chernoff Bound* (Ahlswede and Winter [2002]) which can be stated as follows: and

18. Theorem (Matrix Chernoff Bound). *Let $X_1, \ldots, X_n$ be i.i.d. $m \times m$ matrix valued random variables with expectation M and with $M - cI \preceq X_i \preceq M + cI$, then*

$$\mathbb{P}[\tfrac{1}{n} \sum X_i \notin M \pm \varepsilon I] \leq m \exp(-\varepsilon^2 n / c^2) \tag{23}$$

(One intuition for this bound is that it turns out that diagonal matrices are the hardest ones, and if the distribution was on diagonal matrices, then we need to use the usual Chernoff bound m times and so lose a factor of m in the probability bound.)

In our case, the distribution of X_i's is the distribution of the matrix corresponding to the polynomial $\langle g, x\rangle^4$ whose largest eigenvalue is $\|g\|^4 = k^2$, and so the RHS becomes $k^2 \exp(-\varepsilon^2 n / k^4)$ and so if

[5] **Hint:** One matrix that represents the polynomiaL $Q(v) = \|v\|_2^4$ is the $n^2 \times n^2$ identity matrix.

$n \gg k^4 \log k$ this will suffice. It turns out that (at considerable pain) one can avoid the $\log k$ factor and get the condition $n \gg k^2$. This completes the proof of Lemma 11.

Analyzing success probability

We can now complete the proof of Theorem 8. Let μ be the degree 4 pseudo-distribution satisfying the conditions of the theorem and let $\{u\}$ be the Gaussian distribution that matches its first two moments. By Lemma 11, $\tilde{\mathbb{E}}_\mu \|Px\|_2^2 \leq 0.001$ with high probability, and since $\|Px\|_2^2$ is a degree-2 polynomial, the Quadratic Sampling Lemma implies that $\mathbb{E}\, u\|Pu\|_2^2 \leq 0.001$. By the same argument, $\mathbb{E}\, u\|u\|_2^2 = 1$.

We can now use standard results on the Gaussian distribution to transfer the expectation statements to probability bounds

1. $\mathbb{P}_u \|u\|_2^2 \leq \frac{1}{2} \leq \frac{5}{6}$
2. $\mathbb{P}_u \|Pu\|_2^2 \geq 0.01 \leq 1/10$.

We defer the proofs of 1. and 2. to later, and first argue why they complete the proof of Theorem 8. By combining 1. and 2., with probability at least $1/15$ the algorithm samples u with $\|u\|_2^2 \geq 1/2$ and $\|Pu\|_2^2 \leq 0.01$. In this case, $\|Pu\|_2^2 \leq 0.02\|u\|_2^2$. We assumed $v_0 \perp v_1, \ldots v_k$, which means we can write

$$\|u\|_2^2 = \langle u, v_0\rangle^2 \|v_0\|_2^2 + \|Pu\|_2^2 = \langle u, v_0\rangle^2 + \|Pu\|_2^2. \tag{24}$$

Since $\|Pu\|_2^2$ makes up only a 0.02 fraction of this mass, $\langle u, v_0\rangle$ must make up the rest, and we get $\langle u, v_0\rangle \geq 0.98\|u\|_2^2$. Scaling u to be unit, we recover a unit vector $u/\|u\|$ with very high correlation with v_0.

The success probability can be amplified since by repeatedly sampling a vector u and testing the ratio of $\|u\|_4$ to $\|u\|_2$. So, all that is left to complete the proof of Theorem 8 is to show the proofs of the statements 1. and 2. above.

Proof of 1.: We start with a standard second-moment concentration inequality, which we prove here for completeness. Let X be a nonnegative random variable and let $\theta > 0$. Then

$$\begin{aligned} &\mathbb{E}\, X \leq \theta + \mathbb{P}\, X \geq \theta\, \mathbb{E}[X|X \geq \theta] \\ &\mathbb{E}\, X^2 \geq \mathbb{P}\, X \geq \theta\, \mathbb{E}[X^2|X \geq \theta] \overset{\text{Jensen}}{\geq} \mathbb{P}\, X \geq \theta\, \mathbb{E}[X^2|X \geq \theta]^2. \end{aligned} \tag{25}$$

Combining the equations by eliminating $\mathbb{E}[X|X \geq 0]$ and rearranging gives

$$\mathbb{P}\, X \geq \theta \geq \frac{\mathbb{E}\, X - \theta^2}{\mathbb{E}\, X^2}. \tag{26}$$

We apply this to the random variable $\|u\|_2^2$ for some θ to be chosen later to get

$$\mathbb{P}\,\|u\|_2^2 \ge \theta \ge \frac{{\mathbb{E}\,\|u\|_2^2 - \theta}^2}{\mathbb{E}\,\|u\|_2^4}. = \frac{(1-\theta)}{\mathbb{E}\,\|u\|_2^4}. \tag{27}$$

We need to upper-bound $\mathbb{E}\,\|u\|_2^4$. We expand

$$\mathbb{E}\,\|u\|_2^4 = \sum_{i,j} \mathbb{E}\,u(i)^2 u(j)^2 \overset{\text{Cauchy-Schwarz}}{\le} \sum_{i,j} \sqrt{\mathbb{E}\,u(i)^4}\sqrt{\mathbb{E}\,u(j)^4} = \left(\sum_i \sqrt{\mathbb{E}\,u(i)^4}\right)^2. \tag{28}$$

For fixed i, let μ_i, σ_i be such that $u(i) \sim N(\mu_i, \sigma_i)$. It is a Wikipedia-able fact that

$$\begin{aligned} \mathbb{E}\,u(i)^2 &= \mu_i^2 + \sigma_i^2 \\ \mathbb{E}\,u(i)^4 &= \mu_i^4 + 6\mu_i^2\sigma_i^2 + 3\sigma_i^4. \end{aligned} \tag{29}$$

Hence,

$$\mathbb{E}\,u(i)^4 = \mathbb{E}\,u(i)^{2^2} + 4\mu_i^2\sigma_i^2 + 2\sigma^4 \le 3\,\mathbb{E}\,u(i)^{2^2} \tag{30}$$

which yields

$$\left(\sum_i \sqrt{\mathbb{E}\,u(i)^4}\right)^2 \le 3\left(\sum_i \mathbb{E}\,u(i)^2\right)^2 = 3. \tag{31}$$

So if we pick $\theta = \frac{1}{2}$ we get $\mathbb{P}\,\|u\|_2^2 \ge \frac{1}{2} \ge \frac{1}{6}$.

Proof of 2: This is straight Markov's inequality.

References

Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *STOC*, pages 307–326. ACM, 2012.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *STOC*, pages 31–40. ACM, 2014.

Aditya Bhaskara and Aravindan Vijayaraghavan. Approximating matrix p-norms. In *SODA*, pages 497–511. SIAM, 2011.

Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Transactions on Information Theory*, 60(8):4999–5006, 2014.