SOS Lecture 4: Finding planted sparse vector and dictionary learning / Boaz Barak

Based on scribed (and greatly expanded) notes by Samuel Hopkins and Jerry Li

Finding Sparse Planted Vector

1 Introduction

In this lecture we will see how the SOS algorithm can be used to solve the following problem: Suppose that $V \subseteq R^n$ is a random k-dimensional linear subspace in which someone "planted" a sparse vector v_0. *Sparse* here means that v_0 has few nonzero coordinates in the standard basis—perhaps ϵn. The goal is to recover v_0 given an arbitrary basis of V. We give a more formal description below.

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 $v_0 + \mathrm{Ker}(A)$, 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:

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

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.)

1.1 Formal description of 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 $\mathrm{Span}\{v_0, v_1, \ldots, v_k\}$. 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 the paper [BKS14] 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.)

1.2 Ratios of Norms

Rather than trying to directly trying to find a sparse vector, we will define some 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 lets us not worry 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 a dense/smooth one. 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 $\|1_S\|_q/\|1_S\|_p = (\epsilon n)^{1/q-1/p}$ which means that if $q > p$, the latter ratio is larger than the former by some power of $1/\epsilon$. 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.

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

$$(\mathbb{E}_i\left[v(i)^q\right])^{1/q} \geq \epsilon^{1/q-1/p}(\mathbb{E}_i\left[v(i)^p\right])^{1/p}.$$

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,

$$\begin{aligned}(\mathbb{E}_i\left[v(i)^p\right]) &= \sum_i w(i)\frac{v(i)^p}{n^{q/p}} \\ &\leq (\sum_i w(i)^{1/(1-p/q)})^{1-p/q}(\sum_i v(i)^q/n)^{p/q} \\ &= \epsilon^{1-p/q}(\mathbb{E}_i\left[v(i)^q\right])^{p/q}.\end{aligned}$$

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 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 $\epsilon - 1/n$-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 values $q > p$ for which we know how to compute $\max_{v \in V} \|v\|_q/\|v\|_p$ (e.g. $q = \infty, p \in \{1, 2\}$; not sure if there are any other examples, see also Bhaskara and Vijayaraghavan (SODA 2011) for a discussion of related questions, though note that they are talking of a slightly different question, $\max_{\|v\|_p=1} \|Av\|_q$ for a linear operator A (which for $p = 2$ can encapsulate our question by taking A to be a generator or projector operator to the subspace V, and also, somewhat confusingly, their roles for p and q are switched, and so they mostly deal with the case $q \leq p$ which is not our focus.)

Demanet and Hand [HD13] and Spielman, Wang, and Wright [SWW13] use the ℓ_∞ versus ℓ_1 proxy for sparsity to attack this problem. This can be efficiently computable by running the n linear programs

$$\begin{aligned}
&\max v_i && \text{subject to} \\
&x_i \geq v_i \\
&x_i \geq -v_i \\
&\sum_i x_i = 1 \\
&v \in \operatorname{Span}\{v_0, v_1, \ldots, v_k\}
\end{aligned}$$

and picking the best optimum.

However, if $k \gg 1$, this will not detect a vector v that is 0.01-sparse.

Exercise 1: 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.

Exercise 2:

1. 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 $x \in \mathbb{R}^{\epsilon n}$ with $\|x\|_2^2 = 1$

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

   See footnote for hint[1]

2. Conclude 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 `http://terrytao.wordpress.com/2007/07/02/open-question-deterministic-uup-matrices/` and a paper of Koiran and Ziyzuas connecting this problem to the planted clique problem. (Although note that, unlike the planted clique problem, even a quasipolynomial time algorithm for this problem would be very interesting.)

In the following, we will use ℓ_4 versus ℓ_2 as our proxy for sparsity. A priori this is 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

[1] **Hint:** This uses concentration of measure. See the papers of Guruswami, Lee and Razoborov and Guruswami, Lee and Wigderson for discussion of this result, its proof, and derandomization.

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 $\epsilon\sqrt{n}$ for some $\epsilon > 0$ (which is much better than $k = O(1)$ needed in the ℓ_∞/ℓ_1 case but $k = \Omega(n)$ achieved in the ℓ_2/ℓ_1 case). On the other hand, we don't know how to compute this ratio either. In fact, [BBH$^+$12] 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. However, as mentioned above, this cannot detect 1/100-sparse vectors if the subspace as dimension $\gg \sqrt{n}$:

Exercise 3: Prove that if $V \subseteq R^n$ has dimension $k > \sqrt{n}$ then there is a vector $v \in V$ such that $\mathbb{E}v_i^4 \geq \frac{k^2}{10n}\left(\mathbb{E}v_i^2\right)^2$.

2 Using SoS to Do Better

2.1 Description of the Algorithm

We need to phrase our problem as one of polynomial optimization. We have already mentioned that we will optimize the ratio $\|v\|_4/\|v\|_2$. To be more specific: on input a basis B for the subspace V we have real variables $v_1, \ldots, v_n$. Our program is

$$\begin{aligned} \max \|v\|_4^4 \qquad & \text{subject to} \\ \|v\|_2^2 = 1 & \\ v \in V & \end{aligned}$$

(Note that the condition $v \in V$ can be expressed as $n - k$ linear equations.)

We run the level-4 SoS algorithm on this program to obtain a pseudodistribution $\{v\}$ with attending pseudoexpectation operator $\tilde{\mathbb{E}}$. We then run the Quadratic Sampling Lemma to obtain a random vector $w \in V$ that matches the second moments of $\{v\}$. The result will then follow from the following result

Lemma 2 (Sparse vector recovery— main lemma). *If the subspace $V = \mathrm{Span}\{v_1, \ldots, v_k\}$ is chosen at random and v_0 is ϵ-sparse for $\epsilon \leq k^2/(100000n)$ then $\tilde{\mathbb{E}}\|Pw\|_2^2 \leq 0.01$ 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 (especially since we can repeat the process and output the sparsest vector w we can find).

Remark Note that the algorithm only looks at the first two moments of the distribution $\{v\}$. So, why did we need $\{u\}$ 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.

2.2 Proof of Main Lemma

The SoS algorithm gives us a pseudodistribution satisfying the constraints

$$\mathcal{E} = \left\{ \|v\|_4^4 = C^4/n, \|v\|_2^2 = 1, v \in \operatorname{Span}\{v_0, \dots, v_k\} \right\}$$

where C is some number so that C^4/n is the value of the solution returned by the level-4 SoS relaxation.

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:

Lemma 3 (random subspaces don't contain ℓ_4 versus ℓ_2 sparse vectors—actual distributions)**.** *If $k \ll \sqrt{n}$, with high probability*

$$\|Pv\|_4^4 \leq 10\|Pv\|_2^4/n \tag{1}$$

for every v.

We will show that Lemma 3 implies our Main Lemma for actual distributions. Namely,

Lemma 4 (an ℓ_4 versus ℓ_2 sparse vector must be correlated with v_0—actual distributions)**.** *If P satisfies (1) 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 $\langle w, v_0\rangle^2 \geq 1 - O(1/C)$.*

(Note that this is indeed equivalent to the main lemma since $\|w\|_2^2 = \langle w, v_0\rangle^2 + \|Pw\|_2^2$.)

Proof of Lemma 4. 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,

$$\|w\|_4 \leq \alpha\|v_0\|_4 + \|Pw\|_4$$

which can be rearranged to

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

But since $\|v_0\|_4 = C/n^{1/4}$, and Lemma 3 $\|Pw\|_4 \leq 2/n^{1/4}$, the RHS is at least $1 - 2/C$. □

3 Pseudo-distribution version and proofs

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

Lemma 5 (random subspaces don't contain ℓ_4 versus ℓ_2 sparse vectors—pseudodistributions)**.** *With high probability*

$$\|Pv\|_4^4 \preceq 10\|Pv\|_2^4/n \tag{2}$$

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

Lemma 6 (an ℓ_4 versus ℓ_2 sparse vector must be correlated with v_0—pseudodistributions)**.** *If P satisfies (2) then for every degree 4 pseudo-distribution $\{x\}$ satisfying $\{\|x\|_2^2 = 1, \|x\|_4^4 = \|v_0\|_4^4 = C^4/n\}$ it holds that $\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^2\right] \geq 1 - O(1/C)$.*

Now we test “Marley’s Hypothesis” by lifting the proof of Lemma 4 to a proof of Lemma 6, using Lemma 5 rather than Lemma 3 to do the heavy lifting. We need to be able to mimic all the steps we used when everything is wrapped in pseudoexpectations. The main interesting step was a use of the triangle inequality.

Lemma 7 (Triangle Inequality for Pseudodistributions)**.** *Let $\{x, y\}$ be a degree-4 pseudodistribution. Then*

$$\tilde{\mathbb{E}}\left[\|x+y\|_4^4\right]^{1/4} \leq \tilde{\mathbb{E}}\left[\|x\|_4^4\right]^{1/4} + \tilde{\mathbb{E}}\left[\|y\|_4^4\right]^{1/4}.$$

Exercise 4: Prove Lemma 7

We note that the following easier bound would be fine for us (and follows from past exercises): if the distribution satisfies the constraint $\|x\|_4^4 \geq \|y\|_4^4$ then

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

Proof of Lemma 6 from Lemma 5. The proof is almost identical to the proof of Lemma 4. Let P satisfy

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

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,

$$\tilde{\mathbb{E}}\left[\|x\|_4^4\right]^{1/4} \leq \tilde{\mathbb{E}}\left[\|\langle x, v_0\rangle v_0\|_4^4\right]^{1/4} + \tilde{\mathbb{E}}\left[\|Px\|_4^4\right] = \frac{C}{n^{1/4}}\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^4\right]^{1/4} + \tilde{\mathbb{E}}\left[\|Px\|_4^4\right]^{1/4}.$$

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

$$\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^4\right]^{1/4} \geq \frac{n^{1/4}}{C}(\tilde{\mathbb{E}}\left[\|x\|_4^4\right]^{1/4} - \tilde{\mathbb{E}}\left[\|Px\|_4^4\right]^{1/4}) = 1 - \frac{n^{1/4}}{C}\tilde{\mathbb{E}}\left[\|Px\|_4^4\right]^{1/4}.$$

Now we use our assumption on P to get

$$\tilde{\mathbb{E}}\left[\|Px\|_4^4\right]^{1/4} \leq 2\frac{\tilde{\mathbb{E}}\left[\|Px\|_2^4\right]^{1/4}}{n^{1/4}}.$$

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

$$\tilde{\mathbb{E}}\left[\|Px\|_2^4\right] \leq \tilde{\mathbb{E}}\left[\|x\|_2^4\right].$$

Putting it together, we get

$$\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^4\right]^{1/4} \geq 1 - \frac{2}{C}\tilde{\mathbb{E}}\left[\|x\|_2^4\right]^{1/4}.$$

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

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

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

$$\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^4\right]^{1/4} \geq 1 - \frac{2}{C}.$$

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

$$\tilde{\mathbb{E}}\left[\langle x, v_0\rangle^2 \|x\|_2^2\right] = \tilde{\mathbb{E}}\left[\langle x, v_0\rangle^2\right].$$

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

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

and we are done. □

4 Proof of Lemma 5

True to form, we would like to start by proving Lemma 3 and then lift the proof to the SoS setting. Lets start with a heuristic argument on why would Lemma 3 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. The dominant term in this probability is the probability that one of those g_i's is at least $Cn^{1/4}$ which happens with $\exp(-C^2\sqrt{n})$ probability. So, if $C^2\sqrt{n} \gg k$, we would be able to 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.

4.1 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 $\mathrm{Span}\{v_1, \ldots, v_k\}$,

$$\|Bv\|_4^4 \leq 10\|v\|_2^4/n \tag{3}$$

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 (3) is satisfied.

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

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

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

$\langle g,v\rangle^4 \le 5\|v\|_2^4$ (**Exercise 5:** verify this), 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 $Q(v) \le 10\|v\|_2^4$.

It turns out that "sufficiently large" in this case means as long as $n \gg k^2$.

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

Exercise 6: Let P, Q be two homogenous n-variate degree 4 polynomials, then $P \preceq Q$ 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 $0 \preceq A$ if $w^\top A w \ge 0$ for all w.)

As a corollary, such a polynomial P satisfies $P \preceq \lambda\|x\|_2^4$ if there exists such a matrix M_P with $\|M_P\| \le \lambda$ where $\|M_P\|$ denotes the spectral norm. (Can you see why?)

This connection suggest using the *Matrix Chernoff Bound* and specifically the following theorem

Theorem 8 (Matrix Chernoff Bound, Ahlswede and Winter). *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*

$$\Pr[\tfrac{1}{n}\sum X_i \notin M \pm \epsilon I] \le m\exp(-\epsilon^2 n/c^2)$$

(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 $\|g\|^2 = k$, and so the RHS becomes $k^2\exp(-\epsilon^2 n/k^2)$ and so if $n \gg k^2\log k$ this will suffice. It turns out that (at considerable pain) one can avoid that $\log k$ factor.

5 Full proof of Lemma ??

These next sections contain a great exposition of the full proof with the right, $k = O(\sqrt{n})$ bound, as heroically written by Samuel Hopkings.

5.1 Lemma 5 Holds in Expectation

As in the heuristic argument, the first step in both proofs is to show that the SoS relation we need holds in expectation. For convenience we now change notation and let x be a typical vector in $\mathrm{Span}\{v_1, \ldots, v_k\}$. That is, for some indeterminates $\alpha_1, \ldots, \alpha_k$, we have $x = \sum \alpha_i v_i$. We want to show

$$\|x\|_4^4 \preceq \frac{10\|x\|_2^4}{n}. \tag{4}$$

where both sides are now polynomials in $\alpha_1, \ldots, \alpha_k$. We mechanically expand both sides to the equivalent formulation

$$\sum_s \sum_{i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(s)v_l(s) \preceq \frac{10}{n}\sum_{s,t,i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(t)v_l(t)$$

Our task in this section is to hit both sides with $\mathbb{E}_{v_1,\ldots,v_k}[\cdot]$ and show

$$\mathbb{E}_{v_1,\ldots,v_k}\left[\sum_s \sum_{i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(s)v_l(s)\right] \preceq \mathbb{E}_{v_1,\ldots,v_k}\left[\frac{10}{n}\sum_{s,t,i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(t)v_l(t)\right]. \tag{5}$$

We need to calculate the expected coefficient of every monomial $\alpha_i\alpha_j\alpha_k\alpha_l$ on both the right-and left-hand sides of (5). This is an unpleasant but not terribly difficult case analyis.

Notational Conventions We need to distinguish between ordered multisets of indices i, j, k, l, which we will denote just like that, and sets of indices which do not have repeated elements (even though in our notation some elements may be listed multiple times), which we denote $\{i, j, k, l\}$. When we want to sum over all pairs i, j we write $\sum_{i,j}$, and if we don't want to double-count, we use $\sum_{i\leq j}$.

Left-Hand Side of (5) For each $\{i, j, k, l\}$ we calculate

$$\sum_{\pi\in S_4(i,j,k,l)} \mathbb{E}\big[\sum_s v_{\pi(i)}(s)v_{\pi(j)}j(s)v_{\pi(k)}(s)v_{\pi(l)}(s)\big]$$

which is the coefficient of $\alpha_i\alpha_j\alpha_k\alpha_l$, where S_4 is the symmetric group on the set $\{i, j, k, l\}$.

First note that each term in the sum is identical, so we may equivalently calculate

$$n|S_4(i, j, k, l)|\,\mathbb{E}\big[v_i(1)v_j(1)v_k(1)v_l(1)\big].$$

If one of $\{i, j, k, l\}$ is unique then this is 0. If the $\{i, j, k, l\}$ has exactly two unique elements, then $\mathbb{E}\big[v_i(1)v_j(1)v_k(1)v_l(1)\big] = \mathbb{E}\big[\gamma^2\gamma_2'\big] = 1$, where $\gamma, \gamma' \sim N(0, 1)$, and $|S_4(i, j, k, l)| = 3$. If $\{i, j, k, l\}$ has just one unique element, then $\mathbb{E}\big[v_i(1)v_j(1)v_k(1)v_l(1)\big] = \mathbb{E}\big[\gamma^4\big] = 3$ and $|S_4(i, j, k, l)| = 1$. In sum, the left-hand side is equal to

$$3n\sum_{i\leq j}\alpha_i^2\alpha_j^2. \tag{6}$$

Right-Hand Side of (5) For each i, j, k, l we calculate

$$\sum_{\pi\in S_4(i,j,k,l)}\ \sum_{s,t}\mathbb{E}\left[v_{\pi(i)}(s)v_{\pi(j)}(s)v_{\pi(k)}(t)v_{\pi(l)}(t)\right].$$

We split it into two sums:

$$\sum_{\pi\in S_4(i,j,k,l)}\ \sum_{s=t}\mathbb{E}\left[v_{\pi(i)}(s)v_{\pi(j)}(s)v_{\pi(k)}(t)v_{\pi(l)}(t)\right] + \sum_{\pi\in S_4(i,j,k,l)}\ \sum_{s\neq t}\mathbb{E}\left[v_{\pi(i)}(s)v_{\pi(j)}(s)v_{\pi(k)}(t)v_{\pi(l)}(t)\right].$$

In the first, we have recovered exactly the left-hand side of (5). In the second:

- If $\{i, j, k, l\}$ has some element appearing just once, then the corresponding terms are all 0, as before.

- If $\{i, j, k, l\}$ contains exactly two unique elements then there are four elements π of $S_4(i, j, k, l)$ which will have $\pi(i) = \pi(j)$ and $\pi(k) = \pi(l)$ and $n^2 - n$ terms in the inner sum, so we get $4n^2 - 4n$

- If $\{i, j, k, l\}$ has just one unique element, we have

  $$\mathbb{E}\left[v_{\pi(i)}(s)v_{\pi(j)}(s)v_{\pi(k)}(t)v_{\pi(l)}(t)\right] = \mathbb{E}\left[\gamma^2\gamma'^2\right] = 1.$$

  and so the corresponding sum over $s \neq t$ contributes $n^2 - n$.

All in all, the right-hand side is equal to

$$\frac{10}{n}\left[3n\sum_{i\le j}\alpha_i^2\alpha_j^2+(n^2-n)\left(\sum_{i<j}4\alpha_i^2\alpha_j^2+\sum_i\alpha_i^4\right)\right]. \tag{7}$$

It's now a straightforward exercise to check that (6) $\preceq$ (7).

5.2 First Proof of Lemma 5, $k = O(n^{1/4})$

We now know that (4) holds in expectation, by which we mean that some polynomial $R(\alpha)$ is a sum of squares. Conceptually, what remains to do is show that R is close to its expectation with high probability. What is the right sense of closeness? We will see in the next section that, to achieve the optimum bound of $k = O(n^{1/2})$, we need to interpret "close" to mean that some matrix derived R's coefficient matrix is close to its expectation in the spectral norm.

However, we can achieve $k = O(n^{1/4})$ with a somewhat cruder argument. Our first observation is that

$$\alpha_i\alpha_j\alpha_k\alpha_l \preceq \alpha_i^2\alpha_j^2+\alpha_k^2\alpha_l^2 \tag{8}$$

$$-\alpha_i\alpha_j\alpha_k\alpha_l \preceq \alpha_i^2\alpha_j^2+\alpha_k^2\alpha_l^2. \tag{9}$$

At a high level, the idea is that so as long as the coefficients of the terms $\alpha_i\alpha_j\alpha_k\alpha_l$ which are not squares do not get too big (they are 0 in expectation) and the coefficients of the dominating terms $\alpha_i^2\alpha_j^2$ and $\alpha_k^2\alpha_l^2$ do not get too small, we can use this relation to preserve SoS-ness.

Now we get a little more formal. Let

$$R(\alpha)=\frac{10}{n}\sum_{s,t,i,j,k,l}\alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(t)v_l(t)-\sum_s\sum_{i,j,k,l}\alpha_i\alpha_j\alpha_k\alpha_l v_i(s)v_j(s)v_k(s)v_l(s).$$

We will be a little fast-and-lose with the constants for the sake of readability. In particular, we don't lose too much if we ignore the permutations π and just treat each permutation individually.

We will charge to the coefficient of $\alpha_i^2\alpha_j^2$ the coefficients of $\alpha_i\alpha_j\alpha_k\alpha_l$ for all indices k, l. As long as for all i, j the coefficient of $\alpha_i^2\alpha_j^2$ stays positive when we subtract off the absolute values of the coefficients we are charging to it, R is SoS by (8) and (9).

Unfortunately, even for this cruder version of the argument we need a concentration inequality whose proof is outside scope of these notes. The following statement is a special case of Theorem 1.10 in [SS12], who refer the reader to [Jan97] for a proof.

Theorem 9. *Consider a degree-q polynomial $f(Y) = f(Y_1, \ldots, Y_m)$ of independent centered Gaussian random variables $Y_1, \ldots, Y_m$. Then*

$$\Pr\left[|f(Y)-\mathbb{E}\left[f(Y)\right]|\ge\lambda\right]\le e^2e^{-\left(\frac{\lambda^2}{AVar[f(Y)]}\right)^{1/q}}$$

where A is a universal constant.

We will apply Theorem 9 with the degree-4 polynomials which are the coefficients in $R(\alpha)$. To apply the theorem we must estimate the variances of the coefficients. Let

$$f_{ijkl}=\frac{10}{n}\sum_{s,t}v_i(s)v_j(s)v_k(t)v_l(t)-\sum_s v_i(s)v_j(s)v_k(s)v_l(s).$$

By independence considerations, $Var[f_{ijkl}] \leq Var[f_{iiii}]$. It is a somewhat involved and unenlightening calculation to check that $Var[f_{iiii}] = O(n)$. [2] We assume it here, and note that in the preceding section we showed that the coefficients of $\alpha_i^2\alpha_j^2$ in R are at least $5n$ in expectation.

We are nearly there—the rest of the analysis is a standard combination of the concentration inequality and a union bound, so we will be hand-wavy and ignore the log factors needed to make things precise.

The probability that the coefficient of $\alpha_i^2\alpha_j^2$ is less than $4n$ is $O(e^{-n^{1/2}})$ by application of Theorem 9. On the other hand, the probability that any of the k^2 coefficients (of $\alpha_i\alpha_j\alpha_k\alpha_l$) being charged to $\alpha_i^2\alpha_j^2$ is greater than $3n/k^2$ in absolute value is, again by application of the theorem, at most e^{-3n/k^4}. As long as $k^4 < n$, we can pick constants and *polylog* factors to complete the proof.

5.3 Second Proof of Lemma 5, $k = O(n^{1/2})$

The arguments in the following section first appear in section 7 of [BBH+12], and are fleshed out in [DS].

The first proof loses something in requiring a particular SoS decomposition of R. The following more delicate argument avoids this by using a single application of a concentration inequality to the entire polynomial at once rather than bounding each coefficient separately.

Matrix Concentration Setup Matrix concentration inequalities are the analogue of Chernoff/Bernstein/Azuma/Hoeffding/etc. bounds when the terms being summed are independent or weakly-dependent random matrices rather than scalars. They bound the distance of the resulting random matrix from its expectation in the spectral norm. For a readable treatment of "elementary" matrix concentration inequalities, see [Tro12] or [Tao12]. Most of these inequalities rely on variables which are bounded; our variables instead have Gaussian tails. This could be dealt with by some truncation business, but instead we will use a higher-tech result which simultaneously handles the Gaussian-ness of the underlying distribution and saves a log factor over more elementary methods.

The following presentation follows [BBH+12], with notation somewhat modified to fit these notes. For background on the ψ_p norm (in particular the ψ_2 case) see [Ver10], especially section 5.2.3 on sub-Gaussian random variables.

The ψ_p norm of a distribution $\{a\}$ on $\mathbb{R}^{k3}$ is the least $C > 0$ so that

$$\max_{w \in \mathbb{R}^k, \|w\|_2 = 1} \mathbb{E}\left[e^{\frac{|\langle w, a\rangle|^p}{k^{p/2} C^P}} \right] \leq 2.$$

(We let it be ∞ if no such C exists.) Observe that for $p = 2$ this is quantifying the "Gaussian-ness" of the one-dimensional marginals of the distribution $\{a\}$. The scale factor of $k^{p/2}$ is not in the usual definition but we want to match the theorem statement in [ALPTJ11].

We also require a bounded-ness condition: that there is a constant $K \geq 1$ so that for independent samples $a_1, \ldots, a_n \sim \{a\}$,

$$\Pr\left[\max_{i \leq n} \|a_i\|_2 \geq K(nk)^{1/4} \right] \leq e^{-\sqrt{k}}.$$

[2] To check this, expand the variance as $\mathbb{E}[\cdot^2] - \mathbb{E}[\cdot]^2$ and count how many terms in each resulting sum cancel between the square-expectation and the expectation-squared. Terms always cancel unless indices match up, violating independence.

[3] We have to use k and n here on purpose—when we apply the theorem, each sample will correspond to a dimension of our ambient space and there will have dimension the same as the dimension of our subspace.

Now we can state the main theorem of [ALPTJ11, ALP$^+$10], as stated in [BBH$^+$12] (modulo a minor adjustment of scale factors).

Theorem 10. *Let $\{a\}$ be a distribution on $\mathbb{R}^k$ so that $\mathbb{E}\left[aa^T\right] = I$, the ψ_1 norm of $\{a\}$ is at most $\psi \geq 0$, and the boundedness condition holds for $\{a\}$ with constant K. Let $a_1, \ldots, a_n$ be independent samples from $\{a\}$. Then for some universal constants $c, C > 0$, with probability at least $1 - 2e^{-c\sqrt{k}}$,*

$$(1-\epsilon)I \preceq \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T \preceq (1+\epsilon)I$$

where I is the identity matrix, $\preceq$ is the PSD ordering, and $\epsilon = C(\psi + K)^2\sqrt{k/n}$.

From Matrix Concentration to Lemma 5: Plan of Attack We recall the correspondence between polynomials and coefficient matrices that makes the SoS algorithm tick in the first place. If we can find matrices M_A and M_B for A, B polynomials so that $x^T M_A x = A(x)$ and $x^T M_B x = B(x)$, and if $M_A \preceq M_B$, then $A \preceq B$.

We will use this method to show that two inequalities each hold with high probability:

$$\frac{1}{3}\mathbb{E}\left[\|x\|_2^4\right] \preceq \|x\|_2^4 \tag{10}$$

$$\|x\|_4^4 \preceq 3\,\mathbb{E}\left[\|x\|_4^4\right]. \tag{11}$$

Since we already know

$$\mathbb{E}\left[\|x\|_4^4\right] \preceq \frac{10}{n}\mathbb{E}\left[\|x\|_2^4\right]$$

this gives us Lemma 5 (modulo a minor adjustment of the constants).

Warm Up: Concentration for $\|x\|_2^2$ We turn now to showing that $\|x\|_2^2$ is rarely too much smaller than its expectation. We will be able to leverage this to show that (10) holds with high probability. We expand $\|x\|_2^2$ as a polynomial in $\alpha_1, \ldots, \alpha_k$:

$$\|x\|_2^2 = \sum_{s,i,j} \alpha_i \alpha_j v_i(s) v_j(s).$$

Let $a_1, \ldots, a_n \in \mathbb{R}^k$ be the rows of the matrix whose columns are $v_1, \ldots, v_k$. Let $\{a\}$ be the distribution of the a_i's. Observe that we can rewrite the previous equation as

$$\|x\|_2^2 = \sum_{s,i,j} \alpha_i \alpha_j a_s(i) a_s(j)$$

which has coefficient matrix $\sum_s a_s a_s^T$. Furthermore, $\mathbb{E}\left[aa^T\right] = I$. Now we need to calculate the ψ_1 norm. The distribution $\{a\}$ is rotationally invariant, so we may take $w = e_1$ to be the first standard basis vector, in which case we need to find C so that

$$\mathbb{E}\left[e^{|a(1)|/k^{1/2}C}\right] \leq 2$$

Mathematica says that 2 is an upper bound.

The last thing to check before applying the matrix concentration theorem is the boundedness condition. The event $\max \|a_i\|_2 \geq K(nk)^{1/4}$ is equivalent to the event $\max \|a_i\|_2^2 \geq K^2(nk)^{1/2}$.

Since $\|a_i\|_2^2$ is a degree-2 polynomial of independent Gaussians, we can use Theorem 9 to show that $K = 1$ suffices if $k \approx \sqrt{n}$.[4]

Now we can apply the concentration theorem with $k = \sqrt{n}$ to get that with probability at least $1 - e^{-c\sqrt{k}}$,

$$(1 - 9C^2 n^{-1/4})I \preceq \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T \preceq (1 + 9C^2 n^{-1/4})I.$$

As soon as n gets big enough, this yields

$$0.99\,\mathbb{E}\left[\|x\|_2^2\right] \preceq \|x\|_2^2.$$

Concentration for $\|x\|_2^4$ We now make some observations:

1. If A, B are SoS polynomials with $A \preceq B$, then $A^2 \preceq B^2$. To see this, write $B^2 - A^2 = (B + A)(B - A)$ and note that the multiplicands in the latter are both SoS by hypothesis, so their product is as well.
2. $\mathbb{E}\left[\|x\|_2^4\right]$ and $\mathbb{E}\left[\|x\|_2^2\right]^2$ differ only by a constant factor. The proof is mechanical. In particular,

$$\mathbb{E}\left[\|x\|_2^4\right] \preceq 3\,\mathbb{E}\left[\|x\|_2\right]^2.$$

3. $\|x\|_2^2 \succeq \mathbb{E}\left[\|x\|_2^2\right] \succeq 0$.

Taken all together, we get $0.1\,\mathbb{E}\left[\|x\|_2^4\right] \preceq \|x\|_2^4$ which, modulo some adjustments to the constants, is (10).

Concentration for $\|x\|_4^4$ It remains only to dispatch (11). This we also do by appeal to our matrix concentration theorem, but we will have to be somewhat more careful in using it, for a couple reasons:

1. We will not initially end up with a distribution $\{a\}$ with $\mathbb{E}\left[aa^T\right] = I$.
2. The calculations to show that the ψ_1 and boundedness conditions hold will be somewhat more arduous.

To handle the first of these, we will start in the same way as before by finding the distribution whose empirical covariance matrix is the coefficient matrix of $\|x\|_4^4$, but then we will have to hit these vectors with the pseudo-inverse of $\mathbb{E}\left[\|x\|_4^4\right]$ to get a distribution which has true covariance matrix I. Then we use the lemma below to get the result we want. To handle the second issue, we shut up and calculate. Be prepared for some Taylor series.

Lemma 11. *Let Σ be a symmetric real PSD matrix, $\Sigma^{1/2}$ its square root, Σ^{-1} its pseudo-inverse, and $\Sigma^{-1/2}$ the square root of its pseudo-inverse. Then*

- $\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I.$
- $\Sigma^{1/2}\Sigma^{-1/2} = I.$

[4] Actually we could take K to be $o(1)$ in this case. The analysis where we have to worry about these things is for $\|x\|_4^4$.

Proof. The proofs of both facts are straightforward applications of the characterization of the pseudoinverse and square roots of diagonal PSD matrices (respectively, take the inverses and square roots of the diagonal entries) plus the fact that a symmetric real matrix can be diagonalized. □

We recall that $\|x\|_4^4$ expands as

$$\|x\|_4^4 = \sum_s \sum_{i,j,k,l} \alpha_i \alpha_j \alpha_k \alpha_l v_i(s) v_j(s) v_k(s) v_l(s).$$

Once again, let the vectors $a_1, \ldots, a_n$ be the rows of the matrix whose columns are the v_i's. Then we can rewrite:

$$\|x\|_4^4 = \sum_s \sum_{i,j,k,l} \alpha_i \alpha_j \alpha_k \alpha_l a_s(i) a_s(j) a_s(k) a_s(l).$$

The coefficient matrix of $\|x\|_4^4$ therefore $\sum_s (a_s \otimes a_s)(a_s \otimes a_s)^T$ where $\otimes$ is the tensor product. Let $\{a\}$ again be the distribution of the a_s's, and let $\Sigma = \mathbb{E}\left[(a_s \otimes a_s)(a_s \otimes a_s)^T\right]$ be the true covariance matrix of $(a_s \otimes a_s)$. Since it is a covariance matrix, Σ is symmetric and PSD, so the results of Lemma 11 apply.

Consider now the distribution $\{b\}$ given by $\Sigma^{-1/2}(a \otimes a)$. By Lemma 11, the true covariance matrix satisfies $\mathbb{E}\left[bb^T\right] = I$. Let $b_1, \ldots, b_s$ be independent samples from $\{b\}$. If we could prove

$$\sum_s b_s b_s^T \preceq 3I$$

with high probability, then by hitting both sides with $\Sigma^{1/2}$ on right and left and again applying Lemma 11 we would have (11). Hence, all that remains to do is calculate ψ_1 norm and the boundedness constant for $\{b\}$ in order to apply the matrix concentration theorem.

We begin with the ψ_1-norm calculation. We need an upper bound on C so that

$$\max_{w \in \mathbb{R}^{k^2}, \|w\|_2 = 1} \mathbb{E}\left[e^{\frac{|\langle w, b \rangle|}{kC}}\right] \leq 2.$$

Making the substitution $u = \Sigma^{-1/2} w$, we need to find C so that for all u with $u^T \Sigma u \leq 1$,

$$\mathbb{E}\left[e^{\frac{|\langle \Sigma^{1/2} u, a \otimes a \rangle|}{kC}}\right] \leq 2.$$

The entires of Σ are all either zeroth, first, or second moments of a standard Gaussian, depending on how many repeated indices are at a particular entry. In particular, $\Sigma_{iiii} = 3$ and $\Sigma_{ijij} = \Sigma_{iijj} = \ldots = 1$, and all other entires are 0. (See our analysis of $\mathbb{E}\left[\|x\|_4^4\right]$.)

The condition $u^T \Sigma u \leq 1$, if we interpret u as a $k \times k$ matrix M, implies that

$$\sum_{ij} M_{ij}^2 + \sum_{ij} M_{ij} M_{ji} + \sum_{ij} M_{ii} M_{jj} \leq 1.$$

(where we have dropped some 3's to 1's, which can only reduce the left-hand side). Because $a \otimes a$, considered as a $k \times k$ matrix, is symmetric, we may assume that M is also symmetric (otherwise take $M'_{ij} = (M_{ij} + M_{ji})/2$ and note that the inner product with $a \otimes a$ is preserved). With the symmetry assumption we get

$$2 \sum_{ij} M_{ij}^2 + \left(\sum_i M_{ii}\right)^2 \leq 1$$

which we wastefully use to get

$$\sum_{ij} M_{ij}^2 + \left(\sum_i M_{ii}\right)^2 \leq 1.$$

Since the trace of a matrix is the sum of its eigenvalues, we get

$$\left(\sum_i \lambda_i\right)^2 + \sum_i \lambda_i^2 \leq 1$$

which gives $\sum_i \lambda_i \leq 1$ and $\sum_i \lambda_i^2 \leq 1$. All this shows that it will now suffice to prove that there is $C = O(1)$ so that for every symmetric $k \times k$ matrix M with $TrM \leq 1$ and $TrM^2 \leq 1$,

$$\mathbb{E}\left[e^{\frac{|a^T Ma|}{kC}}\right] \leq 2.$$

By rotational invariance of a, we may actually assume that M is diagonal. Then

$$\frac{1}{kC}|a^T Ma| = \frac{1}{kC}|\sum_i \lambda_i a_i^2| \leq \frac{1}{kC}\left(\left|\sum_i \lambda_i\right| + \left|\sum_i \lambda_i(a_i^2-1)\right|\right) \leq \frac{1}{kC} + \frac{1}{kC}\left|\sum_i \lambda_i(a_i^2-1)\right|$$

where the second-to-last step is the triangle inequality and the last step is since $TrM \leq 1$.

Conditioned on $|\sum_i \lambda_i(a_i^2-1)| \leq 0$, the expectation we're bounding becomes at most $e^{1/kC}+1$, so clearly we can take $C = O(1)$ in this case. So we may assume that at least half of the total expectation comes from the case when $|\sum_i \lambda_i(a_i^2-1)| \geq 0$. By independence and the preceding analysis, we get

$$\mathbb{E}\left[e^{\frac{|a^T Ma|}{kC}}\right] \leq 2e^{1/kC}\prod_i \mathbb{E}\left[e^{\lambda_i(a_i^2-1)/kC}\right]. \tag{12}$$

Now we Taylor expand:

$$\mathbb{E}\left[e^{\lambda_i(a_i^2-1)/kC}\right] = \sum_p \frac{1}{p!}\mathbb{E}\left[\frac{1}{(kC)^p}\lambda_i^p(a_i^2-1)^p\right].$$

We want to bound each term in the Taylor expansion with something involving λ_i^2 so we can use the condition $TrM^2 \leq 1$. Recall that the moments $\mathbb{E}\left[a_i^{2p}\right]$ grow like $\prod_{q \text{ odd, } q \leq p} q \approx p!2^p$. So as long as $1/(kC) \leq 1/16$ or so, recalling that $\sum_i \lambda_i^2 \leq 1$ and therefore $|\lambda_i| \leq 1$ for all i, we can estimate

$$\frac{1}{(kC)^p}\lambda_i^p a_i^{2p} \leq p!(1/kC)^2\lambda_i^2 2^{-p}.$$

Plugging this into the Taylor expansion, using Jensen's inequality on $(a_i^2-1)^p$, and splitting off the terms with $p \leq 2$ gives

$$\mathbb{E}\left[e^{\lambda_i(a_i^2-1)/kC}\right] \leq 2 + (kC)^{-2}\lambda_i^2\sum_{k\geq 2} 2^{-k} \leq e^{O((kC)^{-2}\lambda_i^2)}.$$

Finally, plugging this back into (12) we get

$$\mathbb{E}\left[e^{\frac{|a^T Ma|}{kC}}\right] \leq 2e^{1/kC}e^{O((kC)^{-2})\sum \lambda_i^2} \leq 2$$

for all k and sufficiently small C. So the ψ_1 norm of $\{b\}$ is $O(1)$.

The very last thing we need to do is check the boundedness condition for $\{b\}$. Note that it will suffice to show that $(a \otimes a)$ satisfies the boundedness condition rather than $\{b\}$ if we can show that the largest nonzero eigenvalue of $\Sigma^{-1/2}$ is $O(1)$, which would follow if we could show that the smallest nonzero eigenvalue of Σ is $\Omega(1)$.

The proof of the boundedness condition for $a \otimes a$ is similar to that for $\{a\}$.

6 Analyzing success probability, proof of the qaudratic sampling lemma

We restate the QSL here:

Lemma 12 (Quadratic Sampling Lemma). *If $\{x\}$ is a degree $d \geq 2$ pseudo distribution, then there exists a Gaussian distribution $\{u\}$ such that $\tilde{\mathbb{E}}\big[P(x)\big] = \mathbb{E}\big[P(u)\big]$ for every polynomial P of degree at most 2. This distribution can be efficiently computed from input $\{x\}$.*

Proof. By shifting we can assume that $\tilde{\mathbb{E}}\big[x_i\big] = 0$ for all i. Since $\{x\}$ is a degree 2 pseudo-distribution, its second moment matrix $M = \tilde{\mathbb{E}}\big[x^{\otimes 2}\big] = \tilde{\mathbb{E}}\big[xx^\top\big]$ is psd. Hence, we can write $M = B^\top B$ where B is a $d \times n$ matrix with columns $b_1, \ldots, b_n$ and so $M_{i,j} = \langle b_i, b_j\rangle$. Choose a random standard Gaussian vector $g = (g_1, \ldots, g_n)$ and let $z_i = \langle b_i, g\rangle$.

Then, for every i, j, we get that

$$\mathbb{E}\big[z_i z_j\big] = \mathbb{E}\big[\langle b_i, g\rangle\langle b_j, g\rangle\big] = \sum_{a,b} b_i(a) g_a b_i(b) g_b = \sum a_i(a) b_j(a) = \langle b_i, b_j\rangle = M_{i,j}$$

using the fact that the Gaussians are independent and so $\mathbb{E}\big[g_a g_b\big]$ equals 0 if $a \neq b$ and equals 1 otherwise. □

Proof of Main Theorem from Main Lemma and Quadratic Sampling Lemma. Let $\{u\}$ be the Gaussian distribution obtained from $\{x\}$ which satisfies $\mathcal{E}$. The Main Lemma says that $\tilde{\mathbb{E}}_x\big[\|Px\|_2^2\big] \leq 0.001$ with high probability, and since $\|Px\|_2^2$ is a degree-2 polynomial, the Quadratic Sampling Lemma then implies that $\mathbb{E}_u\big[\|Pu\|_2^2\big] \leq 0.001$. By the same argument, $\mathbb{E}_u\big[\|u\|_2^2\big] = 1$.

We can argue using standard techniques to transfer the expectation statements to probability bounds (the proof comes at the end of the proof of the main theorem).

1. $\Pr_u\left[\|u\|_2^2 \leq \frac{1}{2}\right] \leq \frac{5}{6}$
2. $\Pr_u\left[\|Pu\|_2^2 \geq 0.01\right] \leq 1/10$.

Hence, 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

$$\|u\|_2^2 = \langle u, v_0\rangle^2 \|v_0\|_2^2 + \|Pu\|_2^2 = \langle u, v_0\rangle^2 + \|Pu\|_2^2.$$

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.

By the first part of the main lemma, to test whether it has succeeded, the algorithm simply checks the ℓ_4-versus-ℓ_2 sparsity of the vector u. To succeed with probability $1 - 1/poly(n)$ it will need to sample about $\log n$ times.

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

$$\begin{aligned} &\mathbb{E}\left[X\right] \leq \theta + \Pr\left[X \geq \theta\right] \mathbb{E}\left[X \mid X \geq \theta\right] \\ &\mathbb{E}\left[X^2\right] \geq \Pr\left[X \geq \theta\right] \mathbb{E}\left[X^2 \;\middle|\; X \geq \theta\right] \overset{\text{Jensen}}{\geq} \Pr\left[X \geq \theta\right] \mathbb{E}\left[X^2 \;\middle|\; X \geq \theta\right]^2. \end{aligned}$$

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

$$\Pr\left[X \geq \theta\right] \geq \frac{\mathbb{E}\left[X - \theta\right]^2}{\mathbb{E}\left[X^2\right]}.$$

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

$$\Pr\left[\|u\|_2^2 \geq \theta\right] \geq \frac{\mathbb{E}\left[\|u\|_2^2 - \theta\right]^2}{\mathbb{E}\left[\|u\|_2^4\right]}. = \frac{(1-\theta)}{\mathbb{E}\left[\|u\|_2^4\right]}.$$

We need to upper-bound $\mathbb{E}\left[\|u\|_2^4\right]$. We expand

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

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

$$\begin{aligned} \mathbb{E}\left[u(i)^2\right] &= \mu_i^2 + \sigma_i^2 \\ \mathbb{E}\left[u(i)^4\right] &= \mu_i^4 + 6\mu_i^2\sigma_i^2 + 3\sigma_i^4. \end{aligned}$$

Hence,

$$\mathbb{E}\left[u(i)^4\right] = \mathbb{E}\left[u(i)^2\right]^2 + 4\mu_i^2\sigma_i^2 + 2\sigma^4 \leq 3\,\mathbb{E}\left[u(i)^2\right]^2$$

which yields

$$\left(\sum_i \sqrt{\mathbb{E}\left[u(i)^4\right]}\right)^2 \leq 3\left(\sum_i \mathbb{E}\left[u(i)^2\right]\right)^2 = 3.$$

So if we pick $\theta = \frac{1}{2}$ we get $\Pr\left[\|u\|_2^2 \geq \frac{1}{2}\right] \geq \frac{1}{6}$. □

Proof of (2). This is straight Markov's inequality. □

□

Dictionary Learning

7 Introduction

The *dictionary learning / sparse coding* problem is defined as follows: there is an unknown $n \times m$ matrix $A = (a_1|\cdots|a_m)$ (think of $m = 10n$). We are given access to many examples of the form

$$y = Ax + e \tag{13}$$

for some distribution $\{x\}$ over sparse vectors and distribution $\{e\}$ over noise vectors with low magnitude.

Our goal is to learn the matrix A, which is called a *dictionary*.

The intuition behind this problem is that natural data elements are sparse when represented in the "right" basis, in which every coordinate corresponds to some meaningful features. For example while natural images are always dense in the pixel basis, they are sparse in other bases such as wavelet bases, where coordinates corresponds to edges etc.. and for this reason these bases are actually much better to work with for image recognition and manipulation. (And the coordinates of such bases are sometimes in a non-linear way to get even more meaningful features that eventually correspond to things such as being a picture of a cat or a picture of my grandmother etc. or at least that's the theory behind deep neural networks.) While we can simply guess some basis such as the Fourier or Wavelet to work with, it is best to learn the right basis directly from the data. Moreover, it seems that in many cases it is actually better to learn an *overcomplete* basis: a set of $m > n$ vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ so that every example from our data is a sparse linear combination the a_k's. (Sometimes just considering the case that the a_m's are a union of two bases, such as the standard and Fourier one, already gives rise to many of the representational advantages and computational challenges.)

Olshausen and Field were the first to define this problem - they used a heuristic to learn such a basis for some natural images, and argued that representing images via such an dictionary is somewhat similar to what is done in the human visual cortex. Since then this problem has been used in a great many applications in computational neuroscience, machine learning, computer vision and image processing. Most of the time people use heuristics without rigorous analysis of running time or correctness. There has been some rigorous work using a method known as "Independent Component Analysis", but that method makes quite strong assumptions on the distribution $\{x\}$ (namely independence). Lately, starting with the Spielman-Wang-Wright paper mentioned earlier, there was a different type of rigorously analyzed algorithms, but they all required the vector x to be *very sparse*— less than $\sqrt{n}$ nonzero coordinates. The SOS method allows recovery in the much denser case where x has up to ϵn nonzero coordinates for some $\epsilon > 0$.

Once again this problem has a similar flavor to the "sparse recovery" problem. In the sparse recovery problem, we know the dictionary A (which is also often assumed to have some nice properties such as being random or satisfying "restricted isometry property") and from a single value $y = Ax$ we need to recover x. In the dictionary learning problem we get many examples but, crucially, we know neither A nor x, which makes it a more challenging problem.

7.1 Model

First, we will ignore the vector e in (13). Morally, the SOS algorithm is naturally robust to noise, and thus these small perturbations change little in the analysis, so we will omit them for simplicity. The simplified problem is already quite interesting.

To allow recovery of A, even in the statistical sense, we need to make some assumptions on the distribution $\{x\}$. These assumptions should capture "sparsity". Most rigorous work assumed a hard sparsity constraint, but we will assume a much softer one (as mentioned above). We also make some additional assumptions that are still strictly weaker than those used by most other works (and incomparable to the others). Nevertheless, trying to find the minimal assumptions needed is a great open problem.

Second, we need to make some assumptions on the distribution $\{x\}$ to allow recovery. It will be convenient for us to assume that d is a power of 2. We also will make the following assumption:

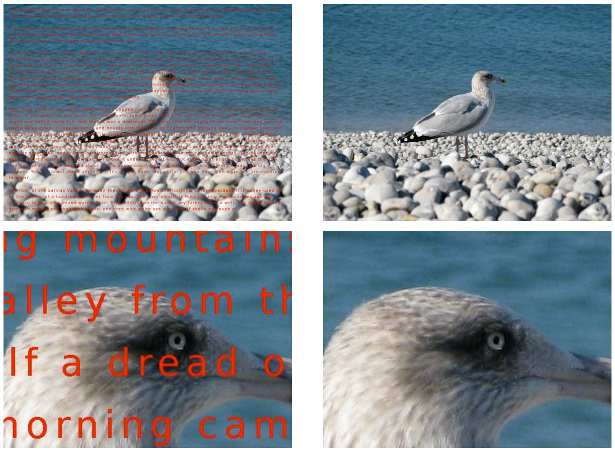

Figure 1: Using dictionary learning to remove overlaid text from images. The authors learned a dictionary A from many natural images, and then removed the text from an image y by (roughly) first representing y as $\sum x_k a^k$ and then zeroing out all the x_i's that are below some threshold. Photos taken from: J. Mairal, F. Bach, J. Ponce, and G. Sapiro. *Online Dictionary Learning for Sparse Coding.* In ICML 2009 (See also Mairal, Julien, Michael Elad, and Guillermo Sapiro. "Sparse representation for color image restoration." , IEEE Transactions on Image Processing 17.1 (2008): 53-69 for a clearer description of the method as well as some nice images of how the dictionary looks like that should be added to the scribe notes...)

for some large constant d, we normalize so that $\mathbb{E}\left[x_i^d\right] = 1$ for every i, and then require that for some parameter $\tau = o(1)$

$$\mathbb{E}\left[x_i^{d/2} x_j^{d/2}\right] \leq \tau \tag{14}$$

for every $i \neq j$. We will also make the additional condition that x_i is somewhat symmetric around zero, in the sense that for every non-square monomial x^α of degree at most d (i.e., $\sum \alpha_i \leq d$ and there is some i for which α_i is odd)

$$\mathbb{E}\left[x^\alpha\right] = 0\ . \tag{15}$$

Condition (14) is essentially minimal, and roughly corresponds to x having at most τn nonzero (or significant) coordinates.

Example 13. For example, note that if the distribution $\{x\}$ is obtain by setting τn random coordinates to equal $\pm\tau^{-(1/d)}$ and the rest zero, then indeed $\mathbb{E}\left[x_i^d\right] = 1$ for all i, and if $i \neq j$

$$\mathbb{E}\left[x_i^{d/2} x_j^{d/2}\right] = \left(\tau\tau^{-(1/2)}\right)^2 = \tau.$$

By appealing to the Arithmetic-Mean-Geometric-Mean inequality, one can show that if assume condition (14) holds with the RHS equalling τ^{4d} (which tends to zero if τ does) then we get the stronger condition

$$\mathbb{E}\left[x^\alpha\right] \leq \tau \tag{16}$$

for every degree d monomial x^α that is not of the form x_i^d. Thus, we call a distribution $\{x\}$ satisfying (16) and (15) (d,τ)-*nice.*

Condition (15) is morally stronger, and it is not clear that it is essential, but it is still fairly natural. In particular for this problem it is without loss of generality to assume that $\mathbb{E}\left[x_i^k\right] = 0$ for every odd k, and so this can be considered a mild generalization of this condition.

We will also assume that every column of A has unit norm, and the spectral norm σ of $AA^\top$ is at most $O(1)$. These are fairly reasonable assumptions as well.

Example 14. For example, consider the case when A is the union of 10 orthonormal bases, so that $m = 10n$. Then for any unit vector $v \in \mathbb{R}^n$, we have that $v^T AA^T v = \|A^T v\|_2 = 10$.

Another minor assumption we make is that $\mathbb{E}\left[x_i^{2d}\right] \leq n^{O(1)}$— this is an extremely mild condition and in some sense necessary for recovery, and so we will not speak much of it except in the one place we use it.

Theorem 15 (Main Result (quasipoly version))**.** *There are some constants $d \in \mathbb{N}, \tau > 0$ and an quasipoly time algorithm R that given poly(n) samples from the distribution $y = Ax$ outputs unit vectors $\{\tilde{a}_1, \ldots, \tilde{a}_m\}$ that are 0.99 close to $\{a_1, \ldots, a_m\}$ in the sense that for every i there is a j such that $\langle a_k, \tilde{a}_j\rangle^2 \geq 0.99$ and vice versa.*

We should note that the paper has a version that runs in polynomial time while requiring sparsity $\tau = n^{-\delta}$ for arbitrarily small $\delta > 0$.

(See the paper for a version that runs in polynomial time while requiring sparsity $\tau = n^{-\delta}$ for arbitrarily small $\delta > 0$.)

Notes on constants:

- In the more general statement the constants d, τ depend on the accuracy (e.g., 0.99) and on the top eigenvalue of $AA^\top$.
- We will think of d as chosen first and then $\tau > 0$ being an extremely small constant depending on d. So for the rest of the analysis we will think of d as some large constant and $\tau = o(1)$.

8 Outline of algorithm

The algorithm is very simple: given examples $y_1, \ldots, y_S$ do the following:

1. Construct the polynomial $\tilde{P}(u) = \frac{1}{S}\sum_{i=1}^{S}\langle y_i, u\rangle^d$
2. Run the SOS algorithm to obtain a degree k pseudo-distribution $\{u\}$ satisfying the constraints $\{\|u\|^2 = 1\}$ that maximizes $\tilde{\mathbb{E}}_u\left[\tilde{P}(u)\right]$. The parameter $k = O(\log n)$ would be specified later.
3. Pick $t = O(\log n)$ random (e.g. Gaussian) vectors $w^1, \ldots, w^t$.
4. Compute the matrix M such that $M_{i,j} = \tilde{\mathbb{E}}\left[\prod_{\ell=1}^{t}\langle w_\ell, u\rangle^2 u_i u_j\right]$.
5. Output a random Gaussian vector v such that $\mathbb{E}\left[v_i v_j\right] = M_{i,j}$.

We will prove the following:

Lemma 16 (Main Lemma)**.** *Suppose the algorithm outputs v. With probability $n^{-O(1)}$, there exists some i such that $\langle v, a_i\rangle^2 \geq 0.99\|v\|^2$.*

The main lemma says that we can get one vector with inverse polynomial probability. We will also show that we can verify when we are successful and so amplify this probability to as close to 1 as we wish. It is unclear how to use a black box reduction to get from this statement recovery of all vectors, but it is possible to do so by a simple extension of the main ideas of this lemma, see the paper for details. Intuitively, all we do at the next step is add a constraint to the SOS algorithm which enforces that the next output we get is far away from the one we've found.

8.1 Proof Outline—Main Ideas

We first give some intuition as to what this algorithm is doing and give an overview of the proof. The first lemma we will need is that $\tilde{P}$ behaves in an interesting way:

Lemma 17. *Let $P(u) = \|A^T u\|_d^d$. For $S \geq ???$ then with probability $\geq ???$*

$$P(u) - \tau\|u\|_d^d \preceq \tilde{P}(u) \preceq P(u) + \tau\|u\|_d^d.$$

Recall that $f \preceq g$ simply means that $g - f$ is a sum of squares.

We make a few straightforward but important observations.

First, notice by repeated usage of the AMGM inequality, if $\{u\}$ satisfies $\tilde{\mathbb{E}}_u\left[\|u\|_2^2\right] = 1$ then it also satisfies $\tilde{\mathbb{E}}_u\left[\|u\|_d^d\right] \leq 1$, thus Lemma 17 implies that

$$\left|\tilde{\mathbb{E}}_u\left[P(u)\right] - \tilde{\mathbb{E}}_u\left[\tilde{P}(u)\right]\right| \leq \tau, \tag{17}$$

so the inequality holds in pseudo-expectation as well.

Second, we see that if u is unit with $P(u) \geq 1$ then it must hold that $\|v\|_d^d \geq 1 - \tau$, but for fixed ϵ, and for τ sufficiently small and d sufficiently large, this implies that there is some i such that $v_k^2 = \langle a^k, u\rangle^2 \geq 1 - \epsilon$. Indeed, otherwise

$$1 - \tau \leq \|v\|_d^d = \sum v_k^d \leq \max_k v_k^{d-2} \sum v_k^2 \leq (1-\epsilon)^d \cdot O(1)$$

and the RHS would be smaller than 1/2 if d is a large enough constant. Importantly, this implies that we have the following:

Corollary 18. *There is an oracle so that given a vector u, returns* Accept *if there exists a_k so that $\langle a_k, u\rangle \geq 1 - O(\tau)$, and* Reject *otherwise.*

Proof. Plug the goddamn thing into $\tilde{P}(u)$. □

Finally, if $\{u\}$ is an actual distribution over unit u's with $P(u) \geq 1$ then every vector in the support would have $\langle a^k, u\rangle^2 \geq 1 - \tau$ for some k. We wish to show that this holds even if it is only a pseudo-distribution. This is captured in the following lemma:

Lemma 19. *Let $\{u\}$ be the pseudo-distribution returned by step 2 of our algorithm. If $t > c \log m$ is an even integer and c sufficiently large then there exists some k_0 such that*

$$\tilde{\mathbb{E}}_u\left[\langle u, a_{k_0}\rangle^t\right] \geq (1-\tau)^{O(t)} . \tag{18}$$

At this point, if $\{u\}$ were a real distribution, then we could just sample from it and we'd be happy, since this lemma would also imply that with some nontrivial probability, $\langle u, a_{k_0}\rangle^2 \geq 1 - O(\epsilon)$. However, it's a pseudo-distribution (boo). The normal thing to do here is just to match the first two moments with a Gaussian, then sample form that. However, if we dropped Step 3-4 and simply tried to define $M_{i,j} = \tilde{\mathbb{E}}\left[u_i u_j\right]$ then sample form a distribution matching these two moments, this will not work:

Example 20. Let us assume that the psuedo-distribution $\{u\}$ was simply the uniform distribution over $\{\pm a_1, \ldots, \pm a_m\}$. This does satisfy all of our conditions. The first moments of this distribution are all zero, and the second moments are $\mathbb{E}\left[u_i u_j\right] = 0$ if $i \neq j$ and $\mathbb{E}\left[u_i^2\right] = 2\sum_k a_{ki}^2$, so what we get is a random linear combination of the a_k. This will not give us any information about the a^k's (in fact can be shown that without loss of generality this would be simply a random vector in $\mathbb{R}^n$, if for instance $a_i = e_i$ where e_i is the ith standard basis vector).

However, the reweighing we do in step 3-4 has the effect that if we are lucky, it will isolate one of the a_k's. To see why in the case of Example 20 note that the matrix M we compute in the particular case above is simply:

$$M = 2\sum f_W(a_k)\cdot (a_k)^{\otimes 2}$$

where for $W = (w_1,\dots,w_t)$, $f_W(a_k) = \prod_{\ell=1}^t \langle w_\ell, a_k\rangle^2$ and for every vector z, $z^{\otimes 2}$ is the matrix Z such that $Z_{i,j} = z_i z_j$.

Intuitively, the idea is that with some inverse polynomial probability, the correlation of each random Gaussian we pick with a_1 will be twice as much as the correlation it has with every other a_k, and hence since we take $O(\log n)$ of these, the weighting will be heavily skewed to $(a_1)^{\otimes 2}$.

In this lucky case we will have that for every i,j $M_{i,j} = f_W(a_1)a_1(i)a_1(j) \pm o(f_W(a_1)/n)$. Therefore, if we sample a random v such that $\mathbb{E}\left[v_i v_j\right] = M_{i,j}$ then (using $\|a_1\| = 1$), we have

$$\mathbb{E}\left[\|v\|^2\right] = \sum_i M_{i,i} = f_W(a_1) \pm o(n f_W(a_1)/n)$$

and

$$\begin{aligned}\mathbb{E}\left[\langle v, a_1\rangle^2\right] &= \sum a_1(i)a_1(j)M_{i,j}\\ &= f_W(a_1)\left(\sum_{i,j}(a_1(i)a_1(j))^2 \pm o(1/n)\sum_{i,j} a_1(i)a_1(j)\right)\\ &= f_W(a_1)(1+o(1/n)\left(\sum a_1(i))^2\right) = f_W(a_1)(1\pm o(1))\end{aligned}$$

Thus, if we scale v to a unit vector $\tilde{v}$, we will get that $\langle \tilde{v}, a_1\rangle^2 \geq 1 - o(1)$.

In general, this behavior is captured in the following lemma:

Lemma 21. *For any degree k pseudo-distribution $\{u\}$ satisfying $\{\|u\|_2^2 = 1\}$ so that there exists some c so that $\tilde{\mathbb{E}}_u\left[\langle u, c\rangle^t\right] \geq e^{-\epsilon k}$, the sampling procedure described in steps 3-5 outputs a c' with $\langle c, c'\rangle \geq 1 - O(\epsilon)$ with probability $2^{-k/\mathrm{poly}(\epsilon)}$.*

By what we've done above, these three lemmas together prove Lemma 16. Since by Corollary 18 we have an oracle to check the correctness of a candidate solution, by repeatedly doing this a polynomial number of times, we conclude that with high probability, we will succeed in finding the desired solution. Now all we have to do is prove a bunch of lemmas, which we do in the remaining sections.

9 Proof of Lemma 17

We restate Lemma 17 here for convenience.

Lemma 22. *Let $P(u) = \|A^T u\|_d^d$. For $S = \mathrm{poly}(\tau, d)$ then with arbitrarily high probability*

$$P(u) - \tau\|u\|_d^d \preceq \tilde{P}(u) \preceq P(u) + 2\tau\|u\|_d^d.$$

Proof. We first show that we can replace $\tilde{P}$ with its expectation. Recall that $\tilde{P}(u) = \frac{1}{S}\sum_{i=1}^S \langle y, u\rangle^d$. Let $Q(u) = \mathbb{E}_y\left[\langle y, u\rangle\right]^d$. Associate to any degree d polynomial $f(x) = \sum_{|\alpha|\leq d} c_\alpha x^\alpha$ a matrix $M(f)$

whose rows and columns are indexed by monomials of degree at most $d/2$ (recall d is even), so that for every monomial β_1, β_2 with $|\beta_1|, |\beta_2| \leq d/2$,

$$M_{\beta_1,\beta_2} = \frac{1}{T_{\beta_1+\beta_2}} c_{\beta_1+\beta_2}$$

where $T_\alpha = \#\{\alpha_1, \alpha_2 : \alpha_1 + \alpha_2 = \alpha\}$. Then it is straightforward to prove that f is a sum of squares if and only if this matrix M is PSD.

We know that $M(\tilde{P}) \to M(Q)$ in the Frobenius norm as we take more and more samples, since all the monomials will converge, and moreover, we know that if we take $\mathrm{poly}(d, \tau)$ samples, we will have that with high probability, $\|M(\tilde{P}) - M(Q)\|_F \leq \tau/2$ and hence in the spectral norm as well, which implies that the matrices $\tau I \pm (M(\tilde{P}) - M(Q))$ are PSD, which by the above is equivalent to the statement that $\pm(\tilde{P} - Q) \preceq \tau\|u\|_d^d$.

We now show that $P(u) \preceq Q(u) \preceq P(u) + \tau\|u\|_d^2$. This combined with what we just proved suffices to complete the proof of Lemma 17. Let us open up this expression for Q. Letting $v = A^\top u$ and recalling that $y = Ax$, we have

$$Q(u) = \mathbb{E}_y\left[\langle y, u\rangle^d\right] = \mathbb{E}_y\left[\langle x, A^T, u\rangle^d\right] = \sum_{|\alpha| \leq d} \mathbb{E}_x\left[x^\alpha v^\alpha\right]$$

noting that the non-square moments here vanish, and that the moments that have more than one variables are at most τ (both by the niceness assumption we place on $\{x\}$), we can see that

$$\|v\|_d^d \preceq Q(u) \preceq \|v\|_d^d + \tau \sum_{|\beta| \leq d/2} v^{2\beta} \preceq \|v\|_d^d + \tau d!(\sum_k v_k^2)^{d/2} \tag{19}$$

where the last inequality follows by repeated application of the AMGM inequality. Note that $\sum_k v_k^2 = \|A^\top u\|_2^2 \preceq O(\|u\|_2^2)$ under our assumption that $\sigma = O(1)$, where σ is the largest singular value of A, so if we choose $\tau \leq O(\frac{1}{d!}) \cdot O(1)^d$ we obtained the desired result.

□

10 Proof of Lemma 19

We restate Lemma 19 here for convenience.

Lemma 23. *Let $\{u\}$ be the pseudo-distribution returned by step 2 of our algorithm. If $t > c \log m$ is divisible by $d - 2$ and c sufficiently large then there exists some k_0 such that*

$$\tilde{\mathbb{E}}_u\left[\langle u, a_{k_0}\rangle^t\right] \geq e^{-\epsilon t/d} \,. \tag{20}$$

where $\epsilon = O(\tau + \log \sigma + d \log \frac{m}{k})$.

We note that if we assume Marley's conjecture, we can prove something much stronger, which is intuitively what we are trying to replicate with this lemma:

Proposition 24. *If the $\{u\}$ returned by our algorithm is a real distribution, then there exists a k_0 so that*

$$\tilde{\mathbb{E}}\left[\langle u, a\rangle^t\right] \geq (1 - \tau)^{\Omega(t)} \geq e^{-\tau \cdot \Omega(t)}.$$

Proof. Since every vector in the support of $\{u\}$ is close to *some* k, there exist k_0 such that with probability at least $1/m$, $\langle u, a\rangle^2 \geq 1 - o(1)$. That means that

$$\tilde{\mathbb{E}}\left[\langle u, a\rangle^t\right] \geq \tfrac{1}{m}(1-\tau)^t \geq (1-\tau)^{t-\log m} \geq 1-\tau)^{\Omega(t)}.$$

□

Proof of Lemma 19. First notice by Lemma 17 we know that there is some $\{u\}$ satisfying $\|u\|_2^2$ with $\tilde{\mathbb{E}}_u\left[\tilde{P}(u)\right] \geq 1 - \tau$; take the real distribution which is identically a_1, for instance. Thus, for the pseudo-expectation that we get, this is satisfied as well, and we get that $\tilde{\mathbb{E}}_u\left[\|A^T u\|_d^d\right] \geq 1 - 2\tau$.

This implies by a straightforward averaging argument as above that there exists some k_0 so that

$$\tilde{\mathbb{E}}_u\left[\langle a_{k_0}, u\rangle^d\right] \geq \frac{1}{m}(1-2\tau).$$

To demonstrate this for larger t, we appeal to the following form of Hölder's inequality:

$$(\|v\|_d^d)^{t/d-2} \preceq (\|v\|_2^2)^{t/(d-2)} \cdot \|v\|_t^t$$

which holds whenever t is an integer multiple of $d-2$. If we substitute in $v = A^T u$, and since moreover $\|A^T u\|_2^2 \preceq \sigma \|u\|_2^2$ where $\sigma = O(1)$ and we assume that our pseudo-expectation satisfies $\{\|u\|_2^2 = 1\}$ we obtain by an additional application of Hölder's inequality

$$\sigma^{t/(d-2)}\,\tilde{\mathbb{E}}_u\left[\|A^T u\|_t^t\right] \geq \tilde{\mathbb{E}}\left[\|v\|_d^d\right]^{t/(d-2)} \geq (1-2\tau)^{t/(d-2)} \geq e^{-2\tau t/(d-2)},$$

so $\tilde{\mathbb{E}}_u\left[\|A^T u\|_t^t\right] = e^{-\Omega(t)}$. We are playing fast and loose with constants a little bit, and the big-Oh here hides some dependencies, but it is all morally correct. Then by the same averaging trick as before, we get the desired result. □

11 Proof of Lemma 21

We again restate the lemma we want to prove here for convenience.

Lemma 25. *For any degree k pseudo-distribution $\{u\}$ satisfying $\{\|u\|_2^2 = 1\}$ so that there exists some c so that $\tilde{\mathbb{E}}_u\left[\langle u, c\rangle^t\right] \geq e^{-\epsilon k}$, the sampling procedure described in steps 3-5 outputs a c' with $\langle c, c'\rangle \geq 1 - O(\epsilon)$ with probability $2^{-k/\mathrm{poly}(\epsilon)}$.*

11.1 Motivation

Here is a crude argument as to why this should happen with good probability in the case described in Example 20, which is roughly what we should expect is the hard case. In this case, for every particular random random vector w, with probability 0.99 that $\max_{i\geq 1}\langle w, a_k\rangle^2 \leq \log m$ but with probability $\exp(-c\log n) = n^{-O(1)}$ we would have that $\langle w, a^k\rangle^2 \geq 2\log m$, and these events are essentially independent if the a^k's are sufficiently close to orthogonal. (In general we can't assume that, but it turns out that this doesn't matter for our final argument.) Hence with $n^{-O(t)} = n^{-O(\log n)}$ probability we would have that for every ℓ and $k > 1$, $\langle w^\ell, a^1\rangle^2 \geq 2\langle w^\ell, a^k\rangle^2$ meaning that for every $k > 1$, $f_W(a^1) \geq 2^t f_W(a^k) = n^2 f_W(a^k)$ if we set $t = 2\log n$.

11.2 Less Heuristic Analysis

First, we need to prove the following technical lemma which says that the sampling procedure is well-defined, since the covariance matrices for Gaussians only make sense if they're PSD. Recall that given $\{u\}$, the matrix that we wish to use to produce our Gaussian is defined as $M_{ij} = \tilde{\mathbb{E}}\left[\prod_{\ell=1}^t \langle w_\ell, u\rangle^2 u_i u_j\right]$. For any choice of $W = (w_1, \ldots, w_t)$ let us define $f_W(u) = \prod_{\ell=1}^t \langle w_\ell, u\rangle^2$.

Lemma 26. *The matrix M is positive semi-definite.*

Proof. For any $x \in \mathbb{R}^n$, we have

$$\begin{aligned} x^T M x &= \sum_{i,j} M_{ij} x_i x_j \\ &= \sum_{ij} \tilde{\mathbb{E}}\left[f_W(u) x_i u_i u_j x_j\right] = \tilde{\mathbb{E}}\left[f_W(u) \sum_{ij} x_i u_i u_j x_j\right] = \tilde{\mathbb{E}}\left[f_W(u)\langle x, u\rangle^2\right] \geq 0 \end{aligned}$$

as f_W is a square polynomial and so is $\langle x, u\rangle^2$. □

We want to prove that with decent (i.e., $n^{-O(1)}$) probability over the choice of the vectors $W = (w_1, \ldots, w_t)$, if we select v that matching the first two moments of

$$\tilde{\mathbb{E}}\left[f_W(u) u^{\otimes 2}\right] \tag{21}$$

then it will satisfy

$$\langle v, a\rangle^2 \geq (1 - O(\epsilon))\|v\|^2 \ . \tag{22}$$

We will prove that (with some decent probability over the choice of W) condition (22) holds in expectation. Since

$$\mathbb{E}\left[\langle v, a\rangle^2\right] = \sum_{ij} \mathbb{E}\left[a_i a_j v_i v_j\right] = \sum_{ij} a_i a_j \, \tilde{\mathbb{E}}\left[f_W(u) u_i u_j\right] = \tilde{\mathbb{E}}\left[f_W(u)\langle u, a\rangle^2\right]$$

and

$$\mathbb{E}\left[\|v\|^2\right] = \sum_{ij} \mathbb{E}_{v_i v_j}\left[=\right] \tilde{\mathbb{E}}\left[f_W(u) \sum_{i,j} u_i u_j\right] = \tilde{\mathbb{E}}\left[f_W(u)\|u\|^2\right]$$

as we choose $\{v\}$ to match the second moments of $\{u\}$, this is equivalent to showing that

$$\tilde{\mathbb{E}}\left[f_W(u)\langle u, a\rangle^2\right] \geq 0.99 \, \tilde{\mathbb{E}}\left[f_W(u)\|u\|^2\right] = 0.99 \, \tilde{\mathbb{E}}\left[f_W(u)\right], \tag{23}$$

where the final equality holds because $\{u\}$ satisfies $\{\|u\|^2 = 1\}$. One needs to add an additional argument to show that this actually happens with decent probability, but it is not a very deep one, and so we skip it here— as always, see the paper for details.

If we select a random standard Gaussian vector w then by the rotation invariance of the Gaussian distribution, $\langle w, c\rangle$ is a standard Gaussian (i.e., distributed per $N(0,1)$), and so $\mathbb{E}\left[\langle w, c\rangle^2\right] = 1$ and the probability that $\langle w, c\rangle^2 \geq 11$ equals some Wikipedia-computable constant $p > 0$.

Let A be this event and let $C \geq 10$ be the the expectation of $\langle w, c\rangle^2 - 1$ conditioned on A.

Note that by the rotation invariance of the Gaussian distribution, $\langle w, b\rangle$ is distributed like $N(0, \|b\|)$ for every $b \perp a$ even after conditioning on A.

For every vector unit u, we can write $u = \langle u, c\rangle c + b$ where $b \perp a$ has norm $\sqrt{1 - \langle u, c\rangle^2}$, and so conditioning on A

$$\mathbb{E}_{w|A}\left[\langle u, w\rangle^2\right] = \langle u, c\rangle^2 \, \mathbb{E}_{w|A}\left[\langle c, w\rangle^2 + 1 - \langle u, c\rangle^2\right] = C\langle c, u\rangle^2 + 1$$

Since $w^1, \ldots, w^t$ are chosen independently, if we condition on A happening for every ℓ (which would occur with probability $p^t = \exp(-O(\log n))$) then, letting $Q(u) = (C\langle u, c\rangle^2 + 1)^t$,

$$\mathbb{E}_{W|A}\big[\tilde{\mathbb{E}}_u\big[f_W(u)\big]\big] = \tilde{\mathbb{E}}_u\big[Q(u)\big]$$

by linearity, and

$$\mathbb{E}_{W|A}\big[\tilde{\mathbb{E}}_u\big[f_W(u)\langle u, c\rangle^2\big]\big] = \tilde{\mathbb{E}}_u\big[Q(u)\langle u, c\rangle^2\big]$$

So, we just need to prove that if $\{u\}$ satisfies our conditions, then

$$\mathbb{E}_u\big[Q(u)\langle u, c\rangle^2\big] \geq 0.99\,\mathbb{E}_u\big[Q(u)\big] \tag{24}$$

We will show that (24) follows from the assumption that $\tilde{\mathbb{E}}_u\big[\langle u, c\rangle^t\big] \geq e^{-\epsilon k}$. Indeed, write $Q(u) = Q'(u) + Q''(u)$ by expanding the expression $Q(u) = (C\langle u, c\rangle^2 + 1)^t$, and letting $Q'(u)$ contains all the terms where we take $C\langle u, c\rangle^2$ to a power larger than $t/2$ and letting $Q''(u)$ contain the rest of the terms.

First, $\mathbb{E}\big[Q''(u)\big]$ is negligible compared to $\mathbb{E}\big[Q(u)\big]$, since the terms in $Q''(u)$ are each of the form $C^s\langle u, c\rangle^{2s}$ for some $s \leq d/2$ and thus since $C^s\langle u, c\rangle^{2s} \preceq C^s\|u\|^{2s}\|c\|^{2s} \preceq C^s\|u\|^{2s}$ and $\{u\}$ satisfies $\{\|u\|^2 = 1\}$, we have that in pseudo-expectation, each of the at most $\binom{t}{t/2}$ terms in $Q''(u)$ is bounded by $(C+1)^{t/2}$, while $Q(u)$ contains the the much larger term $C^t\,\mathbb{E}\big[\langle u, c\rangle^{2t}\big] \geq 0.999^{2t}C^t$.

Thus we can assume $Q(u) = Q'(u)$, but then

$$\tilde{\mathbb{E}}\big[Q'(u)\langle u, a\rangle^2\big] \geq (1 - O(\epsilon))\,\tilde{\mathbb{E}}\big[Q'(u)\big]$$

since we can show this ratio holds for every term of $Q'(u)$ since for every $k \geq t$

$$\tilde{\mathbb{E}}\big[\langle u, a\rangle^{k+2}\big] \geq \tilde{\mathbb{E}}\big[\langle u, a\rangle^k\big]^{(k+2)/k} = \tilde{\mathbb{E}}\big[\langle u, a\rangle^k\big]\,\tilde{\mathbb{E}}\big[\langle u, a\rangle^k\big]^{2/k} \geq (1 - O(\epsilon))\,\tilde{\mathbb{E}}\big[\langle u, a\rangle^k\big]$$

where the first inequality uses Hölder's inequality for psuedo-expectations and the last uses our assumption (20).

This concludes the proof of the Main Lemma for actual distributions.

References

[ALP+10] Radosław Adamczak, Alexander E Litvak, Alain Pajor, Nicole Tomczak-Jaegermann, et al. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *J. Amer. Math. Soc*, 23(2):535–561, 2010.

[ALPTJ11] Radosław Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3):195–200, 2011.

[BBH+12] Boaz Barak, Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 307–326, New York, NY, USA, 2012. ACM.

[BKS14] Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 31–40, New York, NY, USA, 2014. ACM.

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. Technical Report TR14-059, Electronic Colloquium on Computational Complexity (ECCC), April 2014.

[DS] Daniel Dadush and David Steurer. personal communication.

[HD13] P. Hand and L. Demanet. Recovering the Sparsest Element in a Subspace. *ArXiv e-prints*, October 2013.

[Jan97] Svante Janson. *Gaussian hilbert spaces*, volume 129. Cambridge university press, 1997.

[SS12] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 437–446. SIAM, 2012.

[SWW13] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

[Tao12] Terence Tao. Topics in random matrix theory, volume 132 of graduate studies in mathematics. *American Mathematical Society*, 25:76, 2012.

[Tro12] Joel A Tropp. User-friendly tools for random matrices: An introduction. Technical report, DTIC Document, 2012.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.