SOS Lecture 3: Finding a planted sparse vector

Boaz Barak

June 30, 2014

Suggested reading This lecture is based on Section 5 of my paper with Kelner and Steurer "Rounding SOS relaxations" `http://arxiv.org/abs/1312.6652` For a possibly cleaner exposition, see pages 17-19 of the survey with Steurer

For some motivation and prior work on the problem itself, see Paul Hand, Laurent Demanet "Recovering the Sparsest Element in a Subspace" `http://arxiv.org/abs/1310.1654`

Problem description 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. The goal is to recover v_0 given an arbitrary basis of V. (See more formal description below.)

Motivation 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.

Contrast with compressed sensing/sparse recovery The problem is similar to the *sparse recovery* (also known as "compressed sensing") task where one is given an affine subspace A that is of the form $v_0 + V$ where v_0 is sparse and V is an (essentially) random linear subspace, and the goal is to recover v_0.

(Typically the problem is described somewhat differently: we have an $m \times n$ matrix A (often chosen at random), we get the value $y = Av_0$ which determines the $k = n - m$ dimensional affine subspace $v_0 + \mathrm{Ker}(A)$, and 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 the linear program $\min_{i=1}^n x_i$ subject to $x_i \geq v_i, x_i \geq -v_i, v \in A$.) But of course if A was a linear subspace but not affine, then this would return the all zero vectors. (Though see below on variants that do make sense for the planted vector problem.)

Formal description of problem We assume that $v_1, \ldots, v_k \in \mathbb{R}^n$ are chosen randomly as standard Gaussian vectors, and v_0 is some arbitrary unit vector with at most ϵn nonzero coordinates. We are given some 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 $\langle v, v_0 \rangle^2 \geq 0.99$. For simplicity lets also assume that v_0 is orthogonal to $v_1, \ldots, v_k$. (This is not really needed but helps simplify some minor calculations.)

Prior work The following variant of L1 minimization was suggested for this problem (Spielman-Wang-Wright, Demanet-Hand): find a solution $v \in V$ that minimizes $\|v\|_1 = \sum |v_i|$ subject to the condition $\max_i v_i \geq 1$. This can be solved by n linear programs (can you see why?).

The problem is that a random k dimensional subspace V will contain a vector v with $\max v_i \geq 1$ but $\sum |v_i| \leq O(\sqrt{k}/n)$, and so if $|v_0| \gg n/\sqrt{k}$, then even though it is would still with high probability be the sparsest nonzero vector in V, it would not be the solution to this program.

Main result There exists $\epsilon > 0$ such that if $|v_0| \leq \epsilon n$ and $k \leq \epsilon\sqrt{n}$ then we can recover the vector.

Algorithm To use the SOS algorithm we need to translate this problem into polynomial equations. Our intuition for doing so would be that sparse vectors v maximize so called "hyper-contractive" ratios $\|v\|_q/\|v\|_p$ for $q > p$. Specifically, we will assume we know $\|v_0\|_4$ and so consider the following set of equations on variables v

$$\mathcal{E} = \{v \in V, \|v\|_2^2 = 1, \|v\|_4^4 = C^4/n\}$$

When v_0 is a unit vector with $\leq \epsilon n$ nonzero coordinates then if the coordinates all have the same value then $\|v_0\|_4^4 = 1/(\epsilon n)$ and it can be shown that this case minimizes the 4-norm (exercise).

Main Lemma Let $\{v\}$ be a pseudo-distribution satisfying $\mathcal{E}$. Then $\tilde{\mathbb{E}}\|Pv\|^2 \leq 0.001$ where P is the projector to the subspace $\mathrm{Span}\{v_1, \ldots, v_k\}$.

From main lemma to algorithm— use the quadratic sampling lemma:

Gaussian Quadratic Sampling Lemma If $\{v\}$ is a degree $d \geq 2$ pseudo distribution, then there exists a Gaussian distribution $\{u\}$ such that $\tilde{\mathbb{E}}P(v) = \mathbb{E}P(u)$ for every polynomial P of degree at most 2.

Proof of Gaussian Quadratic Sampling Lemma By shifting we can assume that $\tilde{\mathbb{E}}v_i = 0$ for all i. Since $\{v\}$ is a degree 2 pseudo-distribution, its second moment matrix $M = \tilde{\mathbb{E}}v^{\otimes 2} = \tilde{\mathbb{E}}vv^\top$ 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

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

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

QSL implies main theorem Let $\{u\}$ be the Gaussian distribution obtained from $\{v\}$, then $\mathbb{E}\|Pu\|_2^2 \leq 0.001$. Using Markov we can argue that if we sample u from $\{u\}$ then with decent probability $\|Pu\|_2^2 \leq 0.01$ and $\|u\|_2^2 \geq 1/2$, which implies $\|Pu\|_2^2 \leq 0.002\|u\|^2$ and so $\langle u, v_0 \rangle^2 \geq 0.99\|u\|^2$.

Note Where did we use the fact that $\{u\}$ is a degree 4 (as opposed to degree 2) pseudo distribution? We didn't— we only use it in the proof of the main lemma. (Note however that the hypothesis of the Main Lemma doesn't even make sense for pseudo distributions of degree smaller than 4.)

Proof of Main Lemma We first prove the main lemma for actual distribution and then use Marley's corollary. The proof has two parts:

Lemma 1 (actual distributions) With high probability

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

for every v.

Lemma 2 (actual distributions) If P satisfies (1) then for every unit vector $v \in V$ with $\|v\|_4 = \|v_0\|_4 = C/n^{1/4}$, $\langle v, v_0 \rangle^2 \geq 1 - O(1/C)$.

Note that Main lemma follows immediately from Lemma 1 and Lemma 2 since in our case $C \gg 1$.

Proof of Lemma 2 (actual distributions) Lemma 2 is actually quite simple:

If $v \in V$ then we can write $v = \alpha v_0 + Pv$, hence using the triangle inequality

$$\|v\|_4 \leq \alpha\|v_0\|_4 + \|Pv\|_4$$

or

$$\alpha \geq 1 - \|Pv\|_4/\|v_0\|^4$$

Plugging in $\|Pv\|_4 \leq 2\|v\|_2^4/n^{1/4}$ and $\|v_0\|_4 = C/n^{1/4}$ we get the result.

Proof of Lemma 1 (actual distributions): An equivalent formulation is that given an orthonormal basis matrix B for $\mathrm{Span}\{v_1, \ldots, v_k\}$,

$$\|Bv\|_4^4 \leq 10\|v\|_2^4/n \tag{2}$$

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 lets 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 (2) is satisfied.

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

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

That means, that we can think of $P(v) = \|Bv\|_4^4$ as the average of n random polynomials each chosen as $n\langle w, v \rangle^4$ where the w has i.i.d $N(0, 1/\sqrt{n})$. Since in expectation $\langle w, v \rangle^4 \leq 5\|v\|_2^4/n$ (exercise), we can see that if n is sufficiently large then $P(v)$ would be very close to its expectation and so have $P(v) \leq \|v\|_2^4/n$.

It turns out that "sufficiently large" in this case means as long as $n \gg k^2$, but the exercises explore the case $n \gg k^4$ which is a much easier argument.

Testing the "Marley Hypothesis" We now need to show that everything follows through even when $\{v\}$ is not an actual distribution. For this we need to phrase Lemmas 1 and 2 in a p.d. way:

Lemma 1 (pseudo-distributions) With high probability

$$\|Pv\|_4^4 \preceq 10\|Pv\|_2^4/n \tag{3}$$

where we denote $P \preceq Q$ if $Q - P$ is a sum of squares.

Lemma 2 (pseudo-distributions) If P satisfies (3) then for every degree 4 pseudo-distribution $\{v\}$ satisfying $\{\|v\|_2^2 = 1, \|v\|_4^4 = \|v_0\|_4^4 = V^4/n\}$ it holds that $\tilde{\mathbb{E}}\langle v, v_0\rangle^2 \geq 1 - O(1/c)$.

Proof of p.d. Lemma 1 It turns out that the proof of (1) actually already implies (3).

Proof of p.d. Lemma 2 The crucial part is to prove an SOS version of the triangle inequality for the 4-norm.

We can still write $v = \langle v_0, v\rangle v_0 + Pv$ and so

$$\|v\|_4^4 = \|\langle v_0, v\rangle v_0 + Pv\|_4^4$$

Therefore, if we can prove that for every pseudo-distribution of degree at least 4 over vectors $\{(u, w)\}$,

$$\left(\tilde{\mathbb{E}}\|u + w\|_4^4\right)^{1/4} \leq \left(\tilde{\mathbb{E}}\|u\|_4^4\right)^{1/4} + \left(\tilde{\mathbb{E}}\|w\|_4^4\right)^{1/4}$$

then we can carry through the analysis as before.

This is true, but let us prove something a little weaker that still suffices: if $\tilde{\mathbb{E}}\|u\|_4^4 \geq \tilde{\mathbb{E}}\|w\|_4^4$ then

$$\left(\tilde{\mathbb{E}}\|u + w\|_4^4\right)^4 \leq \tilde{\mathbb{E}}\|u\|_4^4 + 15\left(\tilde{\mathbb{E}}\|u\|_4^4\right)^{3/4}\left(\tilde{\mathbb{E}}\|w\|_4^4\right)^{1/4}$$

this will follow from the Cauchy-Schwarz and Holder inequalities, which have SOS proofs.