SOS Lecture 6: Dictionary Learning Via Sum of Squares

Boaz Barak

July 4, 2014

Suggested reading This lecture is based on my paper with Kelner and Steurer ("Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method"). See also Section 4.2 (pages 19-21) in the survey with Steurer.

Problem description 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

$$y = Ax + e \tag{1}$$

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*.

Motivation *[Warning: the following discussion is based on my very rudimentary understanding of this problem area... don't place too much trust in it. Also, while the problem is practical our solution is most certainly not— we need to run the SOS algorithm on a large number of variables and degree k that is a large constant or sometimes even logarithmic. However, one can hope that the ideas behind the algorithm and its analysis could turn out to be useful in getting more efficient algorithms.]* The intuition behind this problem is that natural data elemenl;ts 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 $a^1, \ldots, a^m$ so that every example from our data is a sparse linear combination the a^k's. (Sometimes just considering the case that the a^k'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.) **Notation note:** Recall that we have m vectors $a^1, \ldots, a^m \in \mathbb{R}^n$. I will try to be consistent and have i, j be indices ranging over $[n]$ and k an index that ranges over $[m]$.

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

Figure 1: Using dictionary learning to remove overlaid text from images. The authors (to my understanding) 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...)

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 (and I think— though not sure —not even proof that it converges to the global optimum, but rather only to a local one). 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 $\epsilon > 0$.

Contrast with sparse recovery 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.

Model and main theorem First, we will ignore the vector e in (1). Our justification is that we will allow even coefficients x that are not truly sparse (i.e., do not have any coordinates that are zero) but rather sparse in a looser sense, that some small fraction of coordinates has much bigger magnitude than the rest.

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. We will make the following assumption: for some large constant d, we normalize so that $\mathbb{E}x_i^d = 1$ for every i, and then require that for some parameter $\tau = o(1)$

$$\mathbb{E}x_i^{d/2}x_j^{d/2} \leq \tau \tag{2}$$

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)

$$\mathbb{E}x^\alpha = 0 \, . \tag{3}$$

(To make sure that (3) doesn't trivialize (2), we require d to be a multiple of 4— in fact it will be convenient for us to assume that d is a power of 2, which we assume everywhere below.)

Condition (2) is essentially minimal, and roughly corresponds to x having at most τn nonzero (or significant) coordinates. 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 $\mathbb{E}x_i^d = 1$ for all i, and if $i \neq j$

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

Condition (3) 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 $\mathbb{E}x_i^k = 0$ 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 $O(1)$. This are fairly reasonable assumptions as well. (For example if $m = 10n$ and A is a union of 10 orthogonal bases then $\sigma = 10$— can you see why?)

(Another minor assumption we make is that $\mathbb{E}x_i^{2d} \leq n^{O(1)}$— 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.)

By appealing to the Arithmetic-Mean-Geometric-Mean inequality, one can show that if assume condition (2) holds with the RHS equalling τ^{4d} (which tends to zero if τ does) then we get the stronger condition

$$\mathbb{E}x^\alpha \leq \tau \tag{4}$$

for every degree d monomial x^α that is not of the form x_i^d. Thus, we call a distribution $\{x\}$ satisfying (4) and (3) (d, τ)-*nice.*

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.

(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 $\tau = o(1)$.

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

1. Construct the polynomial $P(u) = \frac{1}{S}\sum_{i=1}^{S} \langle y_i, u\rangle^d$
2. Run the SOS algorithm to obtain a degree k pseudo-distribution $\{u\}$ satisfying the constraints $\{\|u\|^2 = 1, P(u) \geq 1\}$ (e.g., the last constraint is replaced by $P(u) = 1 + y^2$ for some additional auxiliary variable y, but lets not worry about this implementation details too much). 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 $M_{i,j} = \tilde{\mathbb{E}} \prod_{\ell=1}^{t} \langle w^\ell, u\rangle^2 u_i u_j$.
5. Output a random Gaussian vector v such that $\mathbb{E} v_i v_j = M_{i,j}$.

We will prove the following:

Main Lemma With probability $n^{-O(1)}$, there exists some i such that $\langle v, a^k\rangle^2 \geq 0.99\|v\|^2$.

The main lemma says that we can get one vector with inverse polynomial probability. It is not hard to show that we can verify when we are successful and so amplify this probability to as close to 1 as we wish. I do not know of 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.

Proof outline —- actual distributions As usual, we will first prove the main lemma assuming that $\{u\}$ is an actual distribution, and then extend this to pseudo distributions. We will give an overview of the proof, making some simplifying assumptions as we go along— again see the paper for the details.

Assumption We will assume that the polynomial P is identical to its expectation. That is, we assume

$$P(u) = \mathbb{E}\langle y, u\rangle^d = \mathbb{E}\langle Ax, u\rangle^d = \mathbb{E}\langle x, A^\top u\rangle^d$$

this is actually not so problematic, since this is a constant degree polynomial and we can take a large enough polynomial number of samples S to get as close convergence to P as we need.

Consequences Let us open up this expression for P: letting $v = A^\top u$

$$\mathbb{P}(u) = \sum_{|\alpha| \leq d} \mathbb{E} x^\alpha v^\alpha$$

noting that the non-square moments here vanish, and that the moments that have more than one variables are at most τ, we can see that

$$\|v\|_d^d \leq P(u) \leq \|v\|_d^d + \tau \sum_{|\beta| \leq d/2} v^{2\beta} \leq \|v\|_d^d + \tau d! (\sum_k v_k^2)^{d/2} \tag{5}$$

Note that $\sum_k v_k^2 = \|A^\top u\|_2^2 \leq O(\|u\|_2^2)$ under our assumption that $\sigma = O(1)$. Thus we see that if u is unit with $P(u) \geq 1$ then (since $\tau = o(1)$) it must hold that $\|v\|_d^d \geq 1 - o(1)$, but this implies that there is some i such that $v_k^2 = \langle a^k, u\rangle^2 \geq 0.999$. Indeed, otherwise

$$1 - o(1) \leq \|v\|_d^d = \sum v_k^d \leq \max_i v_k^{d-2} \sum v_k^2 \leq 0.999^d O(1)$$

and the RHS would be smaller than $1/2$ if d is a large enough constant.

Bottom line If $\{u\}$ is an actual distribution over unit u's with $P(u) \geq 1$ then every vector in the support would have $\langle a^k, u\rangle^2 \geq 1 - o(1)$ for some k.

Useful corollary Let us record the following corollary of what we proved:

COROLLARY: If $\{u\}$ if $t > c \log m$ is an even integer and c sufficiently large then there exists some k_0 such that for $a = a^{k_0}$,

$$\mathbb{E}_u \langle u, a\rangle^t \geq 0.999^t \ . \tag{6}$$

(Note that by convexity this also implies $\mathbb{E}_u \langle u, a\rangle^k \geq 0.999^k$ for every even $k \geq t$.)

PROOF OF COROLLARY: 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

$$\mathbb{E}\langle u, a\rangle^t \geq \tfrac{1}{m}(1 - o(1))^t \geq (1 - o(1))^{t - \log m} \geq 0.999^t$$

What's next? We just showed that every vector in the support in the distribution $\{u\}$ would be a good solution. Now we just need to argue that Step 5 of the algorithm outputs something close to the support.

This is not immediate. In particular if we dropped Step 3 and simply tried to define $M_{i,j} = \tilde{\mathbb{E}} u_i u_j$ then this will not work.

Heuristic analysis Let us assume that the distribution $\{u\}$ was simply the uniform distribution over $\{\pm a^1, \ldots, \pm a^m\}$. This does satisfy all of our conditions. However, if we sample a Gaussian $\{v\}$ that matches the first two moments of $\{u\}$ we simply get a random linear combination (with Gaussian coefficients) of $a^1, \ldots, a^m$ (can you see why?). 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 $\mathbb{R}^n$).

[Note to potential scribes - in class I showed an explicit example that if the a^k's are the Fourier basis then the unweighted matrix M above is simply $(1/n)$ times the identity and hence gives no information about the basis. Also showed some figures how reweighing affects the probability distribution and changes it from uniform to focused on one vector. Johan Håstad suggested to use the t^{th} power of one Gaussian instead of t independent ones, which might slightly simplify things, or at least notation.]

However, the reweighing has the effect that if we are lucky, it will isolate one of the a^k's. To see why 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 $W = (w^1, \ldots, w^t)$, $f_W(a^k) = \prod_{\ell=1}^t \langle w^\ell, a^k\rangle^2$ and for every vector z, $z^{\otimes 2}$ is the matrix Z such that $Z_{i,j} z_i z_j$.

Now we claim that with probability $n^{-O(1)}$, we will be lucky and (essentially) it will be the case that $f_W(a^1) \gg n \cdot f_W(a^k)$ for every $k > 1$. Lets see a crude argument why this should happen with $n^{-O(\log(n))}$ probability. Indeed, 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 $t = 2 \log n$.

In this lucky case the matrix M will have the form

$$M = f_W(a^1)(a^1)^{\otimes 2} + M'$$

where M' is a matrix where all its entries are bounded by $o(f_W(a^1)/n)$ and hence for every i, j $M_{i,j} = f_W(a^1) a^1_i a^1_j \pm o(f_W(a^1)/n^2)$.

Therefore, if we sample a random v such that $\mathbb{E} v_i v_j = M_{i,j}$ then (using $\|a^1\| = 1$)

$$\mathbb{E}\|v\|^2 = \sum_i M_{i,i} = f_W(a^1) \pm o(n f_W(a^1)/n)$$

and

$$\mathbb{E}\langle v, a^1 \rangle^2 = \sum a^1_i a^1_j M_{i,j} = f_W(a^1)(\sum_{i,j} (a^1_i a^1_j)^2 \pm o(1/n) \sum_{i,j} a^1_i a^1_j = f_W(a^1)(1 + o(1/n)(\sum a^1_i)^2) = f_W(a^1)(1 \pm o$$

(since $\sum_i a^1_i \leq \sqrt{n} \|a^1\|$)

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

Less heuristic analysis Now let's try to make things more concrete. Let $a = a^{k_0}$ for the value k_0 obtained in the corollary above.

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

$$\tilde{\mathbb{E}} f_W(u) u^{\otimes 2} \tag{7}$$

then it will satisfy

$$\langle v, u \rangle^2 \geq 0.99 \|v\|^2 \, . \tag{8}$$

We will prove that (with some decent probability over the choice of W) the conditions (7) and (8) hold in expectation, which amounts to

$$\tilde{\mathbb{E}} f_W(u) \langle u, a \rangle^2 \geq 0.99 \tilde{\mathbb{E}} f_W(u) \|u\|^2 = 0.99 \tilde{\mathbb{E}} f_W(u) \tag{9}$$

(can you see why?). 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, $\langle w, a \rangle$ is a standard Gaussian (i.e., distributed per $N(0, 1)$), and so

$\mathbb{E}\langle w,a\rangle^2 = 1$ and the probability that $\langle w,a\rangle^2 \geq 11$ equals some Wikipedia-computable constant $p > 0$.

Let A be this event and let $C \geq 10$ be the the expectation of $\langle w,a\rangle^2 - 1$ 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 $u = \langle u,a\rangle a + b$ where $b \perp a$ has norm $\sqrt{1-\langle u,a\rangle^2}$, and so conditioning on A

$$\mathbb{E}_{w|A}\langle u,w\rangle^2 = \langle u,a\rangle^2 \mathbb{E}_{w|A}\langle a,w\rangle^2 + 1 - \langle u,a\rangle^2 = C\langle a,u\rangle^2 + 1$$

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 $Q(u) = (C\langle u,a\rangle^2+1)^t$,

$$\mathbb{E}_{W|A,u} f_W(u) = \mathbb{E}_u Q(u)$$

and

$$\mathbb{E}_{W|A,u} f_W(u)\langle u,a\rangle^2 = \mathbb{E}_u Q(u)\langle u,a\rangle^2$$

So, we just need to prove that if $\{u\}$ satisfies our conditions, then there exists k such that for $a = a^k$,

$$\mathbb{E}_u Q(u)\langle u,a\rangle^2 \geq 0.99\mathbb{E}_u Q(u) \tag{10}$$

We will show that (10) reduces to the corollary we proved before, namely that

$$\mathbb{E}\langle u,a\rangle^t \geq 0.999^t$$

Proof of Main Lemma from corollary Indeed, suppose that (6) holds and write $Q(u) = Q'(u) + Q''(u)$ by expanding the expression $Q(u) = (C\langle u,a\rangle^2+1)^t$, and letting where $Q'(u)$ contains all the terms where we take $C\langle u,a\rangle^2$ to a power larger than $t/2$ and letting $Q''(u)$ contain the rest of the terms.

First, $\mathbb{E}Q''(u)$ is negligible compared to $\mathbb{E}Q(u)$, since every 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 $C^t\mathbb{E}\langle u,a\rangle^{2t} \geq 0.999^{2t}C^t$.

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

$$\mathbb{E}Q'(u)\langle u,a\rangle^2 \geq 0.999\mathbb{E}Q'(u)$$

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

$$\mathbb{E}\langle u,a\rangle^{k+2} \geq (\mathbb{E}\langle u,a\rangle^k)^{(k+2)/k} = \mathbb{E}\langle u,a\rangle^k(\mathbb{E}\langle u,a\rangle^k)^{2/k} \geq 0.999\mathbb{E}\langle u,a\rangle^k$$

where the first inequality uses convexity and the last uses our assumption (6).

This concludes the proof of the Main Lemma for actual distributions, given the claim.

Extending to pseudo distributions All the arguments we used in the proof fall in the SOS framework. We briefly illustrate why it is the case. First, our assumption that P is equal to the expectation polynomial is justified since P converges to its expectation also in the spectral norm, and hence would closely approximate it on pseudo expectations as well.

If we go over our justification for the inequality (5), we would see that in fact we can replace every $\leq$ sign there with $\preceq$ in the sense what we really showed was that there are some SOS polynomials S, S' such that

$$\|v\|_d^d + S = P(u) = \|v\|_d^d + \tau d! \|v\|_2^d - S'$$

The corollary can be phrased as using the inequality

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

which has an SOS proof whenever these numbers are all integers.

Finally, all our arguments in justifying (9) used convexity and Cauchy-Schwarz/Holder type arguments that have SOS proofs and hence hold for pseudo expectations as well.