1 Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method

In these notes we consider the algorithm for the dictionary learning problem as described by Barak, Kelner, and Steurer [BKS15]. This algorithm uses a sum-of-squares procedure for solving noisy tensor decomposition as its main tool. This algorithm is an improvement on prior results as it places much fewer requirements on the samples used for learning.

1.1 Preliminaries: Dictionary Learning

In *dictionary learning*, the goal is to learn the columns of a matrix A when given only examples of the form $y = Ax + e$ where x is a sparce vector and e is a noise vector of small magnitude.

This problem has machine learning applications. For example, Mairal, Elad, and Sapiro [MES08] used dictionary learning for image processing: the samples y were images, and the dictionary A was components of these images.

We will consider a version of the dictionary learning problem where A is a *σ-dictionary* (Definition 1) and the vectors x are sampled from a *(d, φ)-nice distribution* (Definition 2). The latter definition allows x to be non-sparse, which makes this a less constrained version of the problem than what has been considered historically.

Definition 1. An $m \times n$ matrix A is a *σ-dictionary* if:

- $\sigma||u||_2^2 - ||A^\top u||_2^2$, when viewed as a polynomial with u as the input, is a sum-of-squares of polynomials.
- All of the columns of A (denoted a^i) are unit vectors.

The first property of a σ-dictionary is equivalent to saying that the *spectral norm* of $A^\top A$ is σ. This can be seen as a proxy for the overcompleteness of A. Here we will consider matrices where $m = O(n)$, so σ will be $O(1)$.

To motivate the definition of (d, φ)-nice distributions (Definition 2), let us first consider the nice properties of sparse vectors that we would like to include in our more relaxed constraints.

First, to assist in calculations we would like the expected value of a certain moment of each coordinate to be normalized to the same value. For simplicity, we will normalize this expected value to 1. Second, we would like each coordinate to be uncorrelated, such that any pair of coordinates has high magnitude with low probability. This second property captures the spirit of sparsity while allowing for distributions where the vectors are not actually sparse.

To motivate the final property, note that without loss of generality $\Pr[x_i = a] = \Pr[x_i = -a]$ in dictionary learning. This is because, if we are given $y = Ax + e$ and $y' = Ax' + e'$, we can consider examples of the form $(y - y')$ instead. This works because $(y - y') = A(x - x') + (e - e')$, where $(x - x')$ is a slightly less "sparse" distribution of vectors and $(e - e')$ is a slightly more noisy distribution of vectors. Indeed, using this trick the expected value of x_i^{2k+1} for any natural number k is 0. We enforce that the expected value of every non-square monomial over the coordinates of x is also 0. This is not without loss of generality, but is a reasonable restriction given the above discussion.

Definition 2. A distribution of n-dimensional vectors is a *(d, φ)-nice distribution* for even d if the following hold:

- $\forall i, \mathbb{E}x_i^d = 1$.
- $\forall$ degree-d monomials $x^\alpha \notin \{x_1^d, x_2^d, \ldots x_n^d\}$, $\mathbb{E}x^\alpha \leq \varphi$.[1]
- $\forall$ degree-d non-square monomials x^α, $\mathbb{E}x^\alpha = 0$.

[1] We write degree-d monomials as x^α, where $\alpha \in [d]^n$, $||\alpha||_1 = d$, and $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$.

The following is a simple example of a (d, φ)-nice distribution.

$$\forall i, x_i = \begin{cases} \varphi^{-\frac{1}{d}} & \text{with probability } \frac{\varphi}{2} \\ -\varphi^{-\frac{1}{d}} & \text{with probability } \frac{\varphi}{2} \\ 0 & \text{otherwise} \end{cases}$$

Note that because our definition of (d, φ)-nice distributions does not enforce strict sparsity, we can consider examples of the form $y = Ax$ by thinking of the small noise vector e as simply a part of x. This will simplify our calculations.

1.2 Overview

We will be able to solve dictionary learning efficiently.

Theorem 3. *For every $\varepsilon > 0$, $\sigma \geq 1$, and $\delta > 0$, there exists a d such that for every σ-dictionary A and $(d, (\varphi = n^{-\delta}))$-nice distribution $\{x\}$, given poly(n) samples from $\{y = Ax\}$, with high probability we can efficiently output a set that is ε-close to the columns of A.*

We define ε-close in the following way.

Definition 4. The *correlation* of two vectors u and v is

$$\mathsf{Cor}(u, v) = \frac{\langle u, v \rangle^2}{||u||_2^2 ||v||_2^2}$$

Definition 5. Two sets S and T of vectors are ε-close if:

- $\forall s \in S$, $\exists t \in T$ such that $\mathsf{Cor}(s, t) \geq 1 - \varepsilon$.
- $\forall t \in T$, $\exists s \in S$ such that $\mathsf{Cor}(s, t) \geq 1 - \varepsilon$.

To achieve the algorithm described in Theorem 3, we use the following result about *Noisy Tensor Decomposition.*

Theorem 6. *For every $\varepsilon > 0$, $\sigma \geq 1$, $\exists d, \tau$ such that for every σ-dictionary A, given a polynomial P such that:*

- *$P(u) - ||A^\top u||_d^d - \tau ||u||_2^d$ is a sum-of-squares of polynomials.*
- *$||A^\top u||_d^d + \tau ||u||_2^d - P(u)$ is a sum-of-squares of polynomials.*

With high probability we can efficiently output a set that is ε-close to the columns of A.

The polynomial P in Theorem 6 is required by the constraints to be "close" to the polynomial $||A^\top u||_d^d$. We will use unit vectors for u, so we get the nice property that $|P(u) - ||A^\top u||_d^d| \leq \tau$.

1.3 Dictionary Learning as Tensor Decomposition

Given the algorithm guaranteed by Theorem 6 as a black box, the only thing we need to do to prove Theorem 3 is to find a polynomial that fits the constraints. Here it is!

$$P(u) = \frac{1}{N} \sum_{i=1}^{N} \langle y_i, u \rangle^d$$

If we have access to enough samples, then this polynomial approaches $\mathbb{E}_y \langle y, u \rangle^d$, which is equal to $\mathbb{E}_x \langle Ax, u \rangle^d$ by the definition of y. In these notes, we will ignore the error introduced by the difference between $P(u)$ and $\mathbb{E}_y \langle y, u \rangle^d$, as it turns out to only affect things by a constant factor.

Lemma 7. *$||A^\top u||_d^d + \varphi\sigma^d d^d ||u||_2^d - \mathbb{E}_x\langle Ax, u\rangle^d$ is a sum-of-squares of polynomials. Furthermore, $\mathbb{E}_x\langle Ax, u\rangle^d - ||A^\top u||_d^d$ is a sum-of-squares of polynomials.*

Lemma 7 shows that $P(u)$ fits the constraints in Theorem 6, with $\tau = \varphi\sigma^d d^d$.

Proof. Consider the following polynomial:

$$p(v) = ||v||_d^d + \varphi d^d ||v||_2^d - \mathbb{E}_x\langle x, v\rangle^d$$

The first part of $p(v)$ ($||v||_d^d$) is the sum of the dth moments of the coordinates of v, that is, $(v_1^d + v_2^d + \dots)$.

The second part of $p(v)$ ($\varphi d^d ||v||_2^d$) is a constant times $(v_1^2 + v_2^2 + \dots)^{d/2}$. This is equal to the sum of all degree-d square monomials.

The third part of $p(v)$ ($\mathbb{E}_x\langle x, v\rangle^d$) is $\mathbb{E}_x(x_1v_1 + x_2v_2 + \dots)^d$.

This third part is subtracted from two parts that are themselves sums-of-squares of polynomials. Therefore, to show that $p(v)$ itself is a sum-of-squares, we simply need to show that things cancel out nicely.

The value $\mathbb{E}_x(x_1v_1 + x_2v_2 + \dots)^d$ is composed of three categories of monomials, which we will analyze using the properties of (d, φ)-nice distributions.

- $\mathbb{E}_x x_i^d v_i^d$. By the first property, we know that each of these is equal to v_i^d, and therefore each of these terms cancels exactly with the corresponding term from the first part of $p(v)$.
- For each square α, we have with multiplicity less than d^d terms of the form $\mathbb{E}_x x^\alpha v^\alpha$. By the second property, these are at most φv^α. Each of these is cancelled out by some term of the second part of $p(v)$, and we have some terms of that second part left over.
- For each non-square α, we have $\mathbb{E}_x x^\alpha v^\alpha = 0$ by the third property, so we don't need to consider this category.

Substitute $v = A^\top u$. Recall that $\sigma^d||u||_2^d - ||A^\top u||_2^d$ is a sum-of-squares of polynomials. Then the upper bound is proven. To show that the lower bound holds, note that $\mathbb{E}_x\langle x, v\rangle^d - ||v||_d^d$ is a sum-of-squares of polynomials by a very similar analysis as above. □

1.4 Noisy Tensor Decomposition

Theorem 6 states that we can find a set of vectors ε-close to A. To find one such vector,

1. Use SOS to find the degree-k pseudodistribution $\{u\}$ that maximizes $P(u)$ s.t. $\|u\|^2 \equiv 1$.
2. Let $W = \prod_{i=1}^{t=O(\log m)} \langle v_i, u\rangle$ where the v_i are standard random Gaussian vectors.
3. Output the top eigenvector of M, where $M_{ij} = \tilde{\mathbb{E}}\, W(u)^2 u_i u_j$.

(To find all the vectors, iterate m times, with the additional requirement that $\langle u, s\rangle < 1 - O(\epsilon)$ for all previous vectors s chosen.) To prove this, first we consider the case where $\{u\}$ is an actual distribution, and then we describe how to generalize the arguments to the case where $\{u\}$ is a pseudo-distribution.

First we show that $\{u\}$ is close to one of the columns of A. If $\{u\}$ is exactly one of the columns of A then

$$P = \|A^T u\|_d^d - \tau\|u\|_2^d \geq 1 - \tau\|u\|_2^d = 1 - \tau$$

But if u is such that $\max_i\langle a_i, u\rangle$ is small then

$$P \approx \|A^T u\|_d^d = \sum_i \langle a_i, u\rangle^d \leq \max_i \langle a_i, u\rangle^{d-2} \sum_i \langle a_i, u\rangle^2 \leq \text{small}^{d-2}\sigma$$

Therefore u is close to one of the a_i. In the worst case (proof: mucking about with the triangle inequality), $\{u\}$ is the uniform distribution on $\{\pm a_1, \dots, \pm a_m\}$. Then,

$$M = \frac{1}{m}\sum_{i=1}^{m} W(a_i)^2 a_i a_i^T$$

If $|W(a_1)| \gg \sqrt{m}|W(a_i)|$ then $M \approx$ (a constant)$a_1 a_1^T$, and the top eigenvector is a_1. We want to show that this happens with probability at least $poly^{-1}(m)$. We know that $W \approx \prod_i \langle v_i, a_j\rangle$ for some fixed column a_j of A. Each $\langle v_i, a_j\rangle$ has expected value 1, and is greater than 2 with probability bounded away from 0, so $\mathbb{P}(W(a_1) > 2^t) \geq \exp(-O(t)) = m^{-O(1)}$. Conditioned on this event, with high probability, for all i, $|W(a^i)| < 1.9^t$, in which case

$$|W(a_1)/W(a_i)| \geq (2/1.9)^t \ll \sqrt{m}$$

Now we need to generalize this to the case where $\{u\}$ is a pseudodistribution. Previously we used $\|v\|_d^d \leq \|v\|_\infty^{d-2}\|v\|_2^2$. Now we need a SOS proof for something like that. First, replace $\|v\|_\infty$ with $\|v\|_k$ (where $k = O(\log m)$ is a multiple of $d-2$), so we want

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

Or for $s = k/(d-2)$,

$$\left(\sum_{i=1}^{m} v_i^d\right)^s \preceq \left(\sum_i v_i^2\right)^s \sum_i v_i^{(d-2)s}$$

By expanding the expressions, we get (where $\binom{s}{\alpha} = \frac{s!}{\alpha_1!\dots\alpha_m!}$)

$$\sum_{|\alpha|=s} \binom{s}{\alpha} v^{d\alpha} \preceq \sum_{|\alpha|=s} \binom{s}{\alpha} v^{2\alpha} \sum_i v_i^{(d-2)s}$$

It suffices to prove $v^{d\alpha} \preceq v^{2\alpha}\sum_i v_i^{(d-2)s}$ for arbitrary α. It suffices to prove $v^{(d-2)\alpha} \preceq \sum_i v_i^{(d-2)s}$. This is implied by the following:

Lemma 8. *Let $w_1, \dots, w_n$ be SOS polynomials. Then $w^\alpha \preceq \sum_i w_i^{|\alpha|}$.*

(Apply this with $w_i = v_i^{d-2}$, since d is even.) The proof is by repeated application of $x \cdot y \preceq \frac{1}{2}x^2 + \frac{1}{2}y^2$. e.g.

$$w_1^3 w_2 = w_1^2 \cdot w_1 w_2 \preceq \frac{1}{2}w_1^4 + \frac{1}{2}w_1^2 \cdot w_2^2 \preceq \frac{1}{2}w_1^4 + \frac{1}{2}(\frac{1}{2}w_1^4 + \frac{1}{2}w_2^4) \preceq w_1^4 + w_2^4$$

(The last step uses the fact that the w_i are SOS.)

The following lemma basically says, "If u is chosen from a (pseudo)-distribution, and $\|A^T u\|_d^d$ is close to 1, then $|\langle c, u\rangle|$ is likely to be close to 1."

Lemma 9. *Let u be a degree-$3k$ pesudodistribution over $\mathbb{R}^n$ such that $\|A^T u\|_d^d \geq e^{-\delta d}$ and $\|u\|_2^2 = 1$. Then there exists a column c of A such that $\tilde{\mathbb{E}}\langle c, u\rangle^k \geq e^{-\epsilon k}$ for $\epsilon = O(\delta + \frac{\log \delta}{d} + \frac{\log m}{d})$.*

Proof. By a SOS version of Holder's Inequality, if $d - 2 \mid k$,

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

Therefore,

$$\begin{aligned}
\|A^T u\|_k^k &\geq \left(\frac{\|A^T u\|_d^d}{\|A^T u\|_2^2}\right)^{k/(d-2)} && \text{SOS Holder's with } v = A^T u \\
&\geq \left(\frac{e^{-\delta d}}{\|A^T u\|_2^2}\right)^{k/(d-2)} && \text{assumption of the lemma} \\
&\geq \left(\frac{e^{-\delta d}}{\sigma \|u\|^2}\right)^{k/(d-2)} && \text{definition of } \sigma \\
&= (e^{-\delta d}/\sigma)^{k/(d-2)} && \|u\|_2 = 1
\end{aligned}$$

There exists a column c of A such that

$$\begin{aligned}
\tilde{\mathbb{E}}\langle c, u\rangle^k &\geq \tilde{\mathbb{E}}\, \|A^T u\|_k^k/m && \text{averaging} \\
&\geq (e^{-\delta d}/\sigma)^{k/(d-2)}/m && \text{the above}
\end{aligned}$$

□

References

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM, 2015.

[MES08] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.