CS 224 Fall 2023
9/18/2023

Scribes: Jonathan Geuter

Lecture 4: Smoothed Analysis and Overcomplete Tensors

We have previously seen how to use Jennrich's algorithm to recover vectors $\{u_i, v_i, w_i\}$ from a tensor

$$T \in \mathbb{R}^{d\times d\times d}, \quad T = \sum_{i=1}^{k} u_i \otimes v_i \otimes w_i \tag{1}$$

given that the $\{u_i\}$ are linearly independent, the $\{v_i\}$ are linearly independent, $d \geq 2$, and no two w_i, w_j are collinear (e.g. a scaled version of the other).[1] However, in the regime $k > d$, Jennrich's algorithm is not guaranteed to recover all vectors anymore. In today's lecture, we are looking at how we can use higher order tensors to handle $k >> d$ by moving beyond worst-case analysis, and instead use *smoothed analysis*.

Let's have a look at one of the examples from previous lectures. We are given samples from a mixture of Gaussians $q = \sum_{i=1}^{k} \lambda_i \mathcal{N}(\mu_i, \mathrm{Id})$, where $\lambda_i \in [0,1]$ and $\sum_i \lambda_i = 1$ (which means to sample from q, we sample from $\mathcal{N}(0, \mathrm{Id})$ with probability λ_i). We have seen how to recover the λ_i and μ_i using various order tensors, such as the third order tensor

$$\mathbb{E}_{x\sim q}[x^{\otimes 3}] = \sum_i \lambda_i \mu_i^{\otimes 3} + \sum_i \lambda_i \mu_i \otimes_3 \mathrm{Id}. \tag{2}$$

However, given sample access to q, we can compute any order tensor given enough compute. For example, we can compute

$$\mathbb{E}_{x\sim q}[x^{\otimes 4}] = \sum_i \lambda_i \mu_i^{\otimes 4} + \sum_i \lambda_i \mu_i^{\otimes 2} \otimes_4 \mathrm{Id} + \sum_i \lambda_i \mathbb{E}[g^{\otimes 4}], \tag{3}$$

where $g \sim \mathcal{N}(0, \mathrm{Id})$. This raises the question if we can somehow leverage Jennrich's algorithm for higher order tensors. A naïve approach would be to contract dimensions, e.g. for

$$T = \sum_i \lambda_i u_i \otimes u_i \otimes u_i \otimes u_i \otimes u_i, \tag{4}$$

[1] This algorithm also works for the more general case $T \in \mathbb{R}^{d_1\times d_2\times d_3}$, in which case the assumption becomes $d_3 \geq 2$.

consider

$$T' = \sum_i \lambda_i \mathrm{vec}(u_i \otimes u_i) \otimes \mathrm{vec}(u_i \otimes u_i) \otimes u_i \tag{5}$$

as an object in $\mathbb{R}^{d^2 \times d^2 \times 2}$, and then apply Jennrich's algorithm hoping that all $\{\mathrm{vec}(u_i \otimes u_i)\}$ are linearly independent. Since $\mathrm{vec}(u_i \otimes u_i)\} \in \mathbb{R}^{d^2}$, one could hope that even for random u_i, these vectors would be linearly independent for $k \sim d^2$. However, it is possible to construct pathological counterexamples like the following. Take $k = 2d$, and $\{u_1, ..., u_{2d}\} = \{a_1, ..., a_d, b_1, ..., b_d\}$ for two orthonormal bases $\{a_i\}$ and $\{b_i\}$. Then

$$\sum_i \mathrm{vec}(a_i \otimes a_i) = \mathrm{Id} = \sum_i \mathrm{vec}(b_i \otimes b_i), \tag{6}$$

which shows that even for k linear in d, the u_i can be linearly dependent. However, in real world applications, the components of a tensor will rarely have a structure as delicate as the union of two orthonormal bases. One approach to overcome such pathological examples is *average-case analysis*, where inputs are completely random. This, on the other hand, oftentimes does not resemble real-world applications either, as inputs usually contain more structure than complete randomness. *Smoothed analysis*, on the other hand, considers taking adversarial examples with added noise, which more closely resembles real-world data and strikes a balance between complete randomness and pure adversarial examples. One can think of the landscape of hardness of a problem, over all possible problem parameters, to have relatively low magnitude in most regions, but have a few narrow spikes, representing adversarial problem initializations. Convolving such parameters with noise smooths out these spikes. Hence, the hope is that with high probability over the added noise, the problem is easy.

Smoothed analysis was introduced by Spielman and Teng in 2001 [ST03], where it was used to show that while the Simplex algorithm provably takes exponential time in certain adversarial cases, its smoothed complexity is polynomial. It has since been applied to many algorithms, such as Gaussian elimination without pivoting [SST05]; various other applications can be found in section 3 of [ST09].

The model we will consider looks as follows: Let $\rho > 0$ be a *smoothing parameter*, k denote the number of components of the tensor, and l be the order of the tensor. Then:

1. Arbitrary vectors $\{u'_{i,j}\}_{i,j}$ for $i \in [k]$, $j \in [l]$ are given
2. $u_{i,j} = u'_{i,j} + \frac{\rho}{\sqrt{d}} g_{i,j}$ for $g_{i,j} \sim \mathcal{N}(0, \mathrm{Id})$ are the smoothed vectors
3. $T = \sum_{i=1}^k u_{i,1} \otimes ... \otimes u_{i,l}$ is observed
4. Try to recover the $u_{i,j}$

Note that we will be dealing with independent perturbations $g_{i,j}$ here. However, oftentimes it makes sense to consider dependent perturbations; consider the case $T = \sum_i u_i^{\otimes l}$. Since for fixed i, all $u_{i,j}$ are equal, we expect them to have the same noise in practice. This setting has e.g. been explored in [BCPV19].

The following theorem leverages smooth analysis to show that with high probability over the randomness of the $\{g_{i,j}\}$, we can recover the $u_{i,j}$ with Jennrich's algorithm:

Theorem 1 ([BCMV14]). *With probability $1 - \frac{1}{\text{superpoly}(d)}$ over the randomness of the $g_{i,j}$, Jennrich's algorithm can recover the $u_{i,j}$, given that*

$$k \leq 0.99 d^{\lfloor \frac{l-1}{2} \rfloor}. \tag{7}$$

Here, superpoly(d) denotes growth faster than polynomial in d. Furthermore, we assume that $\rho \sim \frac{1}{\text{poly}(d)}$, which is a common assumption made in practice. We will not prove the theorem in full. However, we will prove a main ingredient needed in the proof. To get an intuition why it's needed, first consider T' as before (for simplicity, we are considering the case $l = 5$ here):

$$T' = \sum_i \lambda_i \text{vec}(u_i \otimes u_i) \otimes \text{vec}(u_i \otimes u_i) \otimes u_i \tag{8}$$

Now for Jennrich's algorithm to be applicable, we would need $\{\text{vec}(u_i \otimes v_i)\}_{i=1\dots k}$ to be linearly independent, as well as $\{\text{vec}(w_i \otimes x_i)\}_{i=1\dots k}$. However, since we are only given noisy vectors, we require them to be robustly linearly independent instead, in the sense of singular values. To this end, define the *Khatri-Rao product* of U and V, which are the matrices with columns equal to u_i and v_i resp., as

$$W = U \odot V := \left[\text{vec}(u_1 \otimes v_1) \dots \text{vec}(u_k \otimes v_k)\right] \in \mathbb{R}^{d^2 \times k}. \tag{9}$$

Then robust linear independence of the $\{\text{vec}(u_i \otimes v_i)\}_{i=1\dots k}$ translates to lower bounding the smallest singular value of W.

Some of the important concepts of the following proof can e.g. also be found in [Vij20].

Theorem 2 ([BCMV14, ADM$^+$18]). *Let $l = 5$.*[2] *We consider the same setting as before; in particular, assume that*

$$k \leq 0.99 d^{\lfloor \frac{l-1}{2} \rfloor} = 0.99 d^2. \tag{10}$$

Then with probability at least $1 - k\exp(-\Omega(d))$ over the smoothing,

$$\sigma_{\min}(U \odot V) \geq \Omega(\rho^2 / d^2). \tag{11}$$

[2] The theorem also holds for the general case, but we only consider the case $l = 5$ for simplicity.

Proof. Throughout the proof, we will oftentimes make statements that are to be understood as "with high probability over the smoothing", which we will not always explicitly mention. The matrices we observe are

$$\tilde{U} := U + \frac{\rho}{\sqrt{\delta}}\mathcal{N}(0,1)^{d\times k}, \tag{12}$$

i.e. every entry in U is perturbed by some Gaussian noise, and likewise

$$\tilde{V} := U + \frac{\rho}{\sqrt{\delta}}\mathcal{N}(0,1)^{d\times k}. \tag{13}$$

Note that for the proof, we change notation, and what was previously U is now $\tilde{U}$ to emphasize that it's a noisy matrix.

Instead of the minimal singular value, we will consider the *leave-on-out distance* as a proxy, which is easier to manage. For a matrix $M \in \mathbb{R}^{n\times k}$, it is defined as

$$l(M) := \min_i \left\|\Pi_i^{\perp} M_i\right\|, \tag{14}$$

where M_i is the i^{th} column of M, and $\Pi_i^{\perp}$ is the projector to the orthogonal complement of the subspace span$(M_1, M_2, ..., M_{i-1}, M_{i+1}, ..., M_k)$. Hence, $l(M)$ is a measure of how far away a given column of M is from the space spanned by all other columns, minimized over all i. The following lemma motivates the use of the leave-one-out distance:

For any matrix $M \in \mathbb{R}^{n\times k}$, it holds $\sigma_{\min}(M) \geq \frac{1}{\sqrt{k}} l(M)$.

Proof. For any vector $u \in \mathbb{R}^k$, we have (with $S :=$ span$(M_2, ..., M_k)$):

$$\begin{aligned}
l(M) &\leq \left\|\Pi_1^{\perp} M_1\right\| && (15)\\
&= \min_{v\in S} \|M_1 - v\| && (16)\\
&= \min_{\lambda} \left\| M_1 - \sum_{j>1} \lambda_j M_j \right\| && (17)\\
&\leq \left\| M_1 + \sum_{j>1} \frac{u_j}{u_1} M_j \right\| && (18)\\
&= \frac{1}{|u_1|} \left\| \sum_{j=1}^{k} u_j M_j \right\| && (19)\\
&= \frac{1}{|u_1|} \|Mu\|\,. && (20)
\end{aligned}$$

Now taking u to be the minimum singular vector of M, then by definition of the minimum singular vector, this equals

$$\frac{\|u\|}{|u_1|}\sigma_{\min}(M), \tag{21}$$

which proves the statement, because we can assume w.l.o.g. that $|u_1| \geq \frac{1}{\sqrt{k}}\|u\|$, as at least one index u_i has to fulfill this inequality.

This means in order to prove the theorem, it suffices to show that for all $i \in [k]$, we have

$$\left\|\Pi_i^{\perp}(\tilde{U} \odot \tilde{V})_i\right\| \geq \sqrt{k}\Omega(\rho^2/d^2), \tag{22}$$

i.e. that the norm is not too small. This seems tricky to prove, as not only the Khatri-Rao product is random (in the noise added to $\tilde{U}$ and $\tilde{V}$), but so is the projector $\Pi_i^{\perp}$ as it depends on $\tilde{U}$ and $\tilde{V}$. Instead of proving this statement, we can prove a stronger statement:

Let $W \subset \mathbb{R}^{d^2}$ be any subspace with dimension at least $0.01d^2$. Show that $\left\|\Pi_W(\tilde{U} \odot \tilde{V})_i\right\|$ is not too small for all i, where Π_W denotes the projection to W.

Note this is indeed a generalization of the previous statement, as we have $k \leq 0.99d^2$, hence $\dim \Pi_i^{\perp} \geq d^2 - (k-1) \geq 0.01d^2$.

Intuitively, the statement makes sense: W is a very large subspace with dimension of the order of d^2, hence, if we take a random vector and project it onto that subspace, there's a good chance that a decent amount of that vector lies in the subspace W.

We will first prove a "baby version" of this statement, namely a similar statement in dimension d instead of d^2: Letting $W \subset \mathbb{R}^d$ be a subspace with $\dim W \geq 0.01d$, and $\tilde{u} = u + \frac{\rho}{\sqrt{d}}g$ for some $u \in \mathbb{R}^d$ and $g \sim \mathcal{N}(0, \mathrm{Id}_d)$, can we show that $\|\Pi_W \tilde{u}\|$ is not too small?

We will prove the statement in two different ways, the first one being more straightforward, but not providing us with the tools needed to prove the statement in d^2 dimensions, while the second prove will be more closely related to the proof in d^2 dimensions.

Proof 1. Let $D = \dim(W) \geq 0.01d$, and $w_i, ..., w_D$ be an orthonormal basis of W. Recall that $\tilde{u} = u + \frac{\rho}{\sqrt{d}}g$. We have that $\{\langle g, w_i\rangle\}_i$ are independent normal variables.

Furthermore,

$$\|\Pi_W \tilde{u}\| = \big\|(\langle \tilde{u}, w_i\rangle, ..., \langle \tilde{u}, w_D\rangle)\big\| \geq \max_j \big|\langle \tilde{u}, w_i\rangle\big| \tag{23}$$

$$= \max_j \left|\langle u, w_i\rangle + \frac{\rho}{\sqrt{d}}\langle g, w_i\rangle\right|. \tag{24}$$

Now *Gaussian anti-concentration* tells us that for $g \sim \mathcal{N}(0,1)$ and any interval I of length t, we have

$$\mathbb{P}[g \in I] \leq \mathcal{O}(t). \tag{25}$$

Since all $\langle \tilde{u}, w_i\rangle$ are independent Gaussians with variance $\frac{\rho^2}{d}$, centered at random points, this means that

$$\mathbb{P}]\left[\big|\langle \tilde{u}, w_j\rangle\big| \leq t\frac{\rho}{\sqrt{d}} \quad \forall j\right] \leq (\mathcal{O}(t))^D = \exp(-\Omega(d)), \tag{26}$$

where we choose t small enough such that the last equality holds.

Proof 2. We construct a *row echelon basis* of W, namely define $w_1 = (1, \star, \star, ...)$, $w_2 = (0, 1, \star, \star, ...)$, ..., $w_D = (0, ..., 0, 1, \star, \star, ...)$, where all $\star$ entries are absolute value bounded by 1, meaning$\|w_i\| \leq \sqrt{d}$ (which we can assume exists w.l.o.g. by changing the coordinate system if needed). We are now going to "reveal" $\langle \tilde{u}, w_j\rangle$ in reverse order, from $j = D$ to $j = 1$, meaning that we will see that even knowing $\langle \tilde{u}, w_j\rangle$ for $j = D, D-1, ..., i+1$, there will be enough randomness "leftover" in $\langle \tilde{u}, w_i\rangle$ to apply Gaussian anti-concentration. This intuitively makes sense, as all w_j for $j = D, ..., i+1$ have zeros in the first i entries, thus telling us nothing about the randomness in the i^{th} coordinate of $\tilde{u}$.

To make this precise, note that

$$\langle \tilde{u}, w_i\rangle = \langle u, w_i\rangle + \frac{\rho}{\sqrt{d}}g_i + \frac{\rho}{\sqrt{d}}\sum_{j>i}(w_i)_j g_j, \tag{27}$$

so by Gaussian anti-concentration we get

$$\mathbb{P}\left[\big|\langle \tilde{u}, w_i\rangle\big| \leq \mathcal{O}(\rho/\sqrt{d}) \middle| \langle \tilde{u}, w_D\rangle, ..., \langle \tilde{u}, w_{i+1}\rangle\right] \leq \mathcal{O}(1) \tag{28}$$

for the conditional probability (here, we set $t = 1$). This gives us

$$\mathbb{P}\left[\big|\langle \tilde{u}, w_i\rangle\big| \leq \mathcal{O}(\rho/\sqrt{d}) \quad \forall i\right] \leq \exp(-\Omega(d)) \tag{29}$$

as before, which finishes the second proof.

Now the question is how to apply this to the original setting, where we have a subspace W of dimension of order d^2, and a Khatri-Rao product of matrices (i.e. a matrix) instead of a vector. We are not going to go into the details, but at a high level, the strategy is to come up with a two-dimensional equivalent of the row-echelon basis, namely define $W^{(i,j)} \in \mathbb{R}^{d\times d}$ by

$$W^{(i,j)} := \begin{bmatrix} 0 & & & \dots & & & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & & & \dots & & & 0 \\ 0 & \dots & 0 & 1 & \star & \dots & \star \\ \vdots & \ddots & \vdots & \star & \star & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \star & \star & \dots & \star \end{bmatrix} \tag{30}$$

where the 1 lies at the $(i,j)^{\text{th}}$ entry. With this basis at hand, the one-dimension argument is iterated twice, meaning for a fixed i, we can look at the vectors $\{W^{(i,j)}\tilde{v}\}_j$ for some vector $\tilde{v}$ and iterate the one-dimensional argument over these vectors to find a vector $v^{(i)}$ that looks similar to a row-echelon vector. Then, we can iterate the one-dimensional argument over $\{\langle \tilde{u}, v^{(i)}\rangle\}_i$ again to show that these dot products are not too small. □

References

[ADM$^+$18] Nima Anari, Constantinos Daskalakis, Wolfgang Maass, Christos H. Papadimitriou, Amin Saberi, and Santosh Vempala. Smoothed analysis of discrete tensor decomposition and assemblies of neurons, 2018.

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions, 2014.

[BCPV19] Aditya Bhaskara, Aidao Chen, Aidan Perreault, and Aravindan Vijayaraghavan. Smoothed analysis in unsupervised learning via decoupling, 2019.

[SST05] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of the condition numbers and growth factors of matrices, 2005.

[ST03] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time, 2003.

[ST09] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis: An attempt to explain the behavior of algorithms in practice, 2009.

[Vij20] Aravindan Vijayaraghavan. Efficient tensor decomposition, 2020.