CS 224 Fall 2023 Scribes: Clara Mohri
September 18, 2023

# Lecture 4: Tensors III: overcomplete tensor decomposition via smoothed analysis

Please use this file as a template for writing scribe notes. Below we provide some instructions on scribing.

## 1 Introduction

In last lecture, we showed that Jennrich's algorithm is not very noise-robust. Furthermore, for a tensor $T = \sum_{i=1}^{k} u_i^{\otimes 3}$, $u_i \in \mathbb{R}^{d\times d\times d}$, Jennrich's requires the assumption that the $u_i$ vectors are linearly independent. This means that $k \leq d$. However, in the last lecture, we saw examples in which $k > d$ (called the "overcomplete setting") but Jennrich's still succeeds, to an extent (when $k$ is very large it eventually fails). Theoretically speaking, Jennrich's algorithm cannot handle cases in which $k > d$.

This lecture examines the case in which $k \gg d$ by using higher-order tensors.

## 2 Example

This example will illustrate that looking at higher-order tensors can allow to access higher-order tensors.

Consider $q = \sum_{i=1}^{k} \lambda_i \mathcal{N}(\mu_i, I)$, where $I$ is the identity matrix and $\mathcal{N}$ refers to the normal distribution. If $x \sim q$ then the following holds (refer to Lecture 2 for details):

$$\mathbb{E}[x^{\otimes 3}] = \sum_{i=1}^{k} \lambda_i \mu_i^{\otimes 3} + (\sum_{i=1}^{k} \lambda_i \mu_i) \otimes_3 I.$$

Here, we can recover $\mu_i$ for all $i$ as long as they are all linearly independent via Jennrich's algorithm.

Suppose we instead consider $\mathbb{E}[x^{\otimes 4}]$:

$$\begin{aligned}\mathbb{E}[x^{\otimes 4}] &= \sum_{i=1}^{k} \lambda_i \mathbb{E}\left[(\mu_i + g)^4\right] && (g \text{ drawn according to } \mathcal{N}(0, I)) \\ &= \sum_{i=1}^{k} \lambda_i \mu_i^{\otimes 4} + \sum_{i=1}^{k} \lambda_i \mu_i^{\otimes 2} \otimes_4 I + \sum_{i=1}^{k} \lambda_i \mathbb{E}[g^{\otimes 4}]\end{aligned}$$

Now, instead of having a third order tensor as the first term, we have a fourth order tensor. Suppose we have $T = \sum_{i=1}^{k} u_i \otimes u_i \otimes u_i \otimes u_i \otimes u_i$. Recall the function vec: $\mathbb{R}^{d \times d} \to \mathbb{R}^{d^2}$ which takes in a matrix and turns it into a vector. Then we may define $T'$ as follows: $T' = \sum_{i=1}^{k} \lambda_i \text{vec}(u_i \otimes u_i) \otimes \text{vec}(u_i \otimes u_i) \otimes u_i$.

$T'$ is a third-order tensor, and one can apply Jennrich's to recover the tensor decomposition as long as it holds that $\{\text{vec}(u_i \otimes u_i)\}_{i=1}^{k}$ is a linearly independent set. From Lecture $2$, we know that this is implied if $\{u_i\}_{i=1}^{k}$ is a linearly independent set. However, we explore whether Jennrich's can be applied to $T'$ even if the set $\{u_i\}_{i=1}^{k}$ is not linearly independent, since $\{\text{vec}(u_i \otimes u_i)\}_{i=1}^{k}$ is a set of dimension $d^2$ dimension.

To summarize this example: We explore whether one can alter the decomposition of a tensor $T$ which is higher-order, using the vec function. We hope that even if the $u_i$ are not linearly independent, using the vec function will create a linearly independent set, meaning that we could use Jennrich's when $k$ is on the order $d^2$.

## 2.1 Counterexample

This counter example shows that we cannot go past $k > O(d)$. Suppose $k = 2d$, and let $\{a_i\}_{i=1}^{d}, \{b_i\}_{i=1}^{d}$ be two orthonormal basis in $\mathbb{R}^d$. Then consider the set of vectors $\{u_i\}_{i=1}^{2d} = \{a_1, \ldots, a_d, b_1, \ldots, b_d\}$. Clearly this is a linearly dependent set of vectors. Furthermore, we make the following claim:

**Claim 1.** *The set $V = \{\text{vec}(u_i \otimes u_i)\}_{i=1}^{2d}$ are linearly dependent.*

*Proof.* Because $\{a_i\}_{i=1}^{d}$ is an orthonormal basis, we know that

$$\sum_{i=1}^{d} a_i a_i^\top = \sum_{i=1}^{d} a_i \otimes a_i = I.$$

Furthermore, the same holds for orthonormal basis $\{b_i\}_{i=1}^{d}$:

$$\sum_{i=1}^{d} b_i \otimes b_i = I.$$

This shows that

$$\begin{aligned} \sum_{i=1}^{d} a_i \otimes a_i = \sum_{i=1}^{d} b_i \otimes b_i &\implies \sum_{i=1}^{d} \text{vec}(a_i \otimes a_i) = \sum_{i=1}^{d} \text{vec}(b_i \otimes b_i) \\ &\implies \exists S \subset [2d], \exists T \subset [2d], S \cap T = \emptyset \text{ s. t. } \sum_{i \in S} V_i = \sum_{i \in T} V_i. \end{aligned}$$

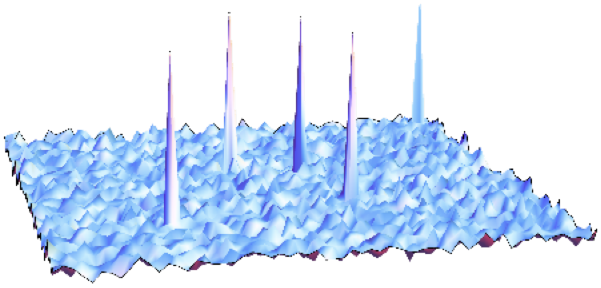

Figure 1: This figure attempts to illustrate the “landscape of difficulty” of a problem over problem instances. The $x, y$ planes simplify the notion of different parameters of the underlying problem instance, and the $z$ axis represents “hardness”.

This proves that the set $V$ is in fact linearly dependent. As such, this is a valid counter example proving that by reshaping using the vec function, we do not obtain the desired hope that one could obtain linear independence and apply Jennrich’s algorithm. □

One should note that in reality, it is not likely that our components are delicately chosen such that they are the union of two orthonormal basis. In the context of Figure 1, this counterexample could be viewed as one of the peaks.

## 3 Beyond Worst-Case Analysis

We introduce the concept of “average case analysis.” An early description of this comes from [Lev86]: “Many interesting combinatorial problems were found to be NP-complete. Since there is little hope to solve them fast in the worst case, researchers look for algorithms which are fast just ‘on average’. This matter is sensitive to the choice of a particular NP-complete problem and a probability distribution of its instances.”

This represents **Perspective 1**: For any given random problem instance, the probability of it being hard is small.

Smoothed analysis is a development on top of average case analysis, and was introduced by [ST03], introduced to analyze typical examples we encounter in reality. This represents **Perspective 2**: For an arbitrary problem instance, if we add to it a small amount of noise, the probability that the noised instance is still hard is small.

# 4 Smoothed Analysis of Tensor Decomposition

Our model is as follows: $\rho > 0$ is the smoothing parameter and dictates how much noise we add to each sample, $k$ is the number of components, and $l$ is the order of the tensor.

Then we receive the problem instance as follows:

1. Nature picks arbitrary vectors $\{u'_{i,j}\}$ for $i \in [k], j \in [l]$.
2. Nature samples $u_{i,j} = u'_{i,j} + \frac{\rho}{\sqrt{d}} \cdot g_{i,j}$ where $g_{i,j} \sim \mathcal{N}(0, I)$.
3. We observe $T = \sum_{i=1}^{k} u_{i,1} \otimes \cdots \otimes u_{i,l}$.

**Theorem 1.** *With prob $1 - 1/\text{superpoly}(d)$ over the randomness of $\{g_{i,j}\}$ we can recover the $u_{i,j}$ for all $i, j$ given $T$, if $k \leq 0.99 d^{\lfloor (l-1)/2 \rfloor}$ [BCMV13].*

Going back to the example, suppose we have some tensor $T' = \sum_i \text{vec}(u_i \otimes v_i) \otimes (w_i \otimes x_i) \otimes y_i)$, we would like to show that $\{\text{vec}(u_i \otimes v_i)\}_{i=1}^{k}$ is robustly linearly independent (due to addition of noise) as well as $\{\text{vec}(w_i \otimes x_i)\}_{i=1}^{k}$.

**Definition 1.** *(Khatri-Rao Product) The Khatri-Rao product of $U$ and $V$ is: $U \star V = \big(\text{vec}(u_1 \otimes v_1) \cdots \text{vec}(u_k \otimes v_k)\big)$ where each entry is a column.*

Thus we would like to show that $U \star V$ is robustly full-rank, in the sense that its minimum value is not too small. Specifically, the following theorem is useful:

**Theorem 2.** *In the smoothed analysis setting, with probability at least $1 - k\exp(-\Omega(d))$: $\sigma_{\min}(U \star V) \geq \text{poly}(\rho, 1/\sqrt{d})$. [BCMV13, ADM+18]*

*Proof.* Begin with $U, V \in \mathbb{R}^{d \times k}$, which are the worst case matrices. Then the smoothed matrices are:

$$\tilde{U} = U + \frac{\rho}{\sqrt{d}} \mathcal{N}(0, I)^{d \otimes k}$$
$$\tilde{V} = V + \frac{\rho}{\sqrt{d}} \mathcal{N}(0, I)^{d \otimes k}$$

The statement that we prove is that $\sigma_{\min}(\tilde{U} \star \tilde{V}) \geq \text{poly}(\rho^2/d^3)$ with high probability. Note that Theorem 2 refers to $U, V$ as worst-case matrices.

We first employ a reduction for minimum singular value to leave-one-out distance:

**Definition 2.** *(Leave-one-out distance) For $M \in \mathbb{R}^{n \times k}$, the leave-one-out distance of $M$, denoted $l(M)$, is defined as $l(M) = \min_{i \in [k]} ||\Pi_i^{\perp} M_i||_2$. $\Pi_i^{\perp}$ is the projector to the orthogonal complement of $\text{span}(M_1, \ldots, M_{i-1}, M_{i+1}, \ldots, M_k)$.*

**Lemma 1.** $\sigma_{\min}(M) \geq \frac{1}{\sqrt{k}} l(M)$.

*Proof.* Take any $i \in [k]$. Without loss of generality, let $i = 1$. Then the following holds:

$$
\begin{aligned}
l(M) &\leq ||\Pi \frac{1}{1} M_1||_2 \\
&= \min_{v \in \text{span}(M_2, \ldots, M_k)} ||M_1 - v||_2 && \text{(by definition)} \\
&= \min_{\vec{\lambda}} ||M_1 - \sum_{i>1} \lambda_i M_i||_2 && \text{(re-writing explicitly)} \\
&\leq ||M_1 + \sum_{i>1} \frac{u_i}{u_1} M_i||_2 && \text{(Taking } \lambda_i = -u_i/u_1\text{)} \\
&= \frac{1}{|u_1|} || \sum_{i=1}^{k} u_i M_i||_2 && \text{(by summing over all } i\text{)} \\
&= \frac{1}{|u_1|} ||Mu||_2 && \text{(by taking } u \text{ to be the minimum singular vector of } M\text{)} \\
&= \frac{||u||_2}{|u_1|} \sigma_{\min}(M) && \text{(property of minimum singular vector)}
\end{aligned}
$$

The largest entry in $u$ is always at least $\frac{1}{\sqrt{k}}||u||_2$. So in general, we can replace $u_1$ in the last line with the largest absolute entry in $u$ to obtain: $l(M) \leq \sqrt{k}\sigma_{\min}(M) \implies \frac{1}{\sqrt{k}} l(M) \leq \sigma_{\min}(M)$. □

Since we have shown that the minimum singular value of a matrix is lower-bounded by a factor times $l(M)$, it suffices to show that $l(\tilde{U} \star \tilde{V})$ is not too small in order to show that $\sigma_{\min}(\tilde{U} \star \tilde{V})$ is not too small. In other words, it suffices to show that $\forall i \in [k], ||\Pi \frac{1}{i} (\tilde{U} \star \tilde{V})_i||_2$ is not too small.

Note that $\dim(\Pi \frac{1}{i}) = d^2 - (k-1)$ in this setting because the matrix is in $d^2$ and there are $k-1$ columns which make the span. By assumption, $k \leq .99d^2$. Hence $\dim(\Pi \frac{1}{i}) \geq .01d^2 = \Omega(d^2)$.

In order to remove some aspect of randomness, we focus on proving a stronger statement: Fix $W \subset \mathbb{R}^{d^2}$ such that $\dim(W) \geq .01d^2$. Then $||\Pi_W(\tilde{U} \star \tilde{V})_i||_2$ is not too small for all $i \in [k]$, with high probability.

We begin with a toy example. Namely, we start with one vector instead of two: Let $W \subset \mathbb{R}^d, \dim(W) \geq .01d$. If $\tilde{u} \in \mathbb{R}^d$ is given by $\tilde{u} = u + \frac{\rho}{\sqrt{k}} g$ ($g \sim \mathcal{N}(0, I)$) then $||\Pi_W \tilde{u}||_2$ is not too small with high probability.

*Proof.* Let $D = \dim(W), D \geq .01d$. Let $\{w_1, \ldots, w_D\}$ be an orthonormal basis for $W$. Then $\langle g, w_1 \rangle, \ldots, \langle g, w_D \rangle$ are all independent random variables (by orthogonality of basis).

$$\begin{aligned}||\Pi_W \tilde{u}||_2 &= ||(\langle w_1, \tilde{u}\rangle \cdots \langle w_D, \tilde{u}\rangle)||_2 && (W \text{ is an orthonormal basis}) \\ &\geq \max_{j\in[D]} \left|\left\langle w_j, \tilde{u}\right\rangle\right| && (\text{ L2 norm is at least value of largest entry})\end{aligned}$$

Every entry $\langle w_i, \tilde{u}\rangle$ can be written as $\langle w_i, u\rangle + \frac{\rho}{\sqrt{d}} \langle w_i, g\rangle$. Recall that all the $\langle w_i, g\rangle$ are independent Gaussian random variables, and because we are taking the dot product of a random Gaussian and a normal vector, the term $\frac{\rho}{\sqrt{d}} \langle w_i, g\rangle$ is in fact a random normal variable with mean $0$ and variance $\frac{\rho^2}{d}$. Thus, all the dot products $\langle w_1, \tilde{u}\rangle, \ldots, \langle w_D, \tilde{u}\rangle$ are independent Gaussians with variance $\frac{\rho^2}{d}$, and they each have some arbitrary mean due to the arbitrary $u$.

**Fact 1.** *(Gaussian anti-concentration). For $g \sim \mathcal{N}(0,1)$ and for any interval $I \subset \mathbb{R}$ of length $t$, $\Pr[g \in I] \leq \Omega(t)$.*

The proof follows from that the probability of landing in interval $I$ is bounded above by some rectangle of width $t$ which contains the area under the curve of the probability density function. By anti-concentration, $\Pr[|\left\langle w_j, \tilde{u}\right\rangle| \leq t\frac{\rho}{\sqrt{d}}] \leq \Omega(t)$. This implies that $\Pr[|\left\langle w_j, \tilde{u}\right\rangle| \leq t\frac{\rho}{\sqrt{d}}, \forall j] \leq \exp(-\Omega(d))$. This shows that with high probability, $||\Pi_W \tilde{u}||_2$ is at $O(\frac{\rho}{\sqrt{d}})$. □

The above proof does not generalize to many vectors. Thus we show a proof that does generalize. We pick a row-echelon basis for $W$ of the form ($\star$ indicates arbitrary value whose absolute value is upper-bounded by 1) and without loss of generality, because it exists up to permutation:

$$\begin{aligned}w_1 &= (1, \star, \ldots, \star) \\ w_2 &= (0, 1, \star, \ldots, \star) \\ w_3 &= (0, 0, 1, \star, \ldots, \star) \\ &\vdots\end{aligned}$$

We reveal $\left\langle w_j, \tilde{u}\right\rangle$ in reverse order $j = D, D-1, \ldots, 1$. That is, we look at $\langle w_i, \tilde{u}\rangle$ conditioned on $\langle w_D, \tilde{u}\rangle, \ldots, \langle w_{i+1}, \tilde{u}\rangle$.

$$\begin{aligned}\langle w_i, \tilde{u}\rangle &= \langle w_i, u\rangle + \frac{\rho}{\sqrt{d}} \langle w_i, g\rangle \\ &= \langle w_i, u\rangle + \frac{\rho}{\sqrt{d}}(g_i(w_i)_i + \sum_{j>i} g_j(w_i)_j)\end{aligned}$$

$w_j$ is arbitrary for $j > i$ so we can think of $\sum_{j>i} g_j(w_i)_j$ as arbitrary. Hence, $g_i(w_i)_i$ is still independent, even when conditioned on $\langle w_{i+1}, \tilde{u} \rangle, \ldots, \langle w_D, \tilde{u} \rangle$. Applying anti-concentration, we have that

$$\Pr\left[\forall i, |\langle w_i, \tilde{u} \rangle| \leq O\left(\frac{\rho}{\sqrt{d}}\right)\right] \leq \exp(-\Omega(d)).$$

In general however, we care about the scenario in which there is $||\Pi_W \tilde{u} \star \tilde{v}||_2$ and $W$ is a subspace of dimension $d^2$. This will require a 2 dimensional row-echelon basis $(W^{(i,j)}) \subset \mathbb{R}^{d \times d}$. We construct $W^{i,j}$ to be a matrix in $\mathbb{R}^{d \times d}$ where

$$(W^{i,j})_{a,b} = \begin{cases} 0, \text{if } a < i \text{ or } b < j \\ 1, \text{if } a = i \text{ and } b = j \\ *, \text{otherwise} \end{cases}$$

High level idea: we can look at the vectors $\{W^{i,j}\tilde{v}\}_j$. For every $i$, we can extract vector $v^i$ and look at $\left\langle v^i, \tilde{u} \right\rangle$. If the $v^i$ are not too different from row-echelon vectors, then we can show that this is not too small.

□

# References

[ADM+18] Nima Anari, Constantinos Daskalakis, Wolfgang Maass, Christos H. Papadimitriou, Amin Saberi, and Santosh S. Vempala. Smoothed analysis of discrete tensor decomposition and assemblies of neurons. *CoRR*, abs/1810.11896, 2018.

[BCMV13] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. *CoRR*, abs/1311.3651, 2013.

[Lev86] Leonid A. Levin. Average case complete problems. *SIAM Journal on Computing*, 15(1):285–286, 1986.

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