3

Tensor Decompositions

Algorithms

In this chapter, we will study tensors and various structural and computational problems we can ask about them. Generally, many problems that are easy over matrices become ill-posed or *NP*-hard when working over tensors instead. Contrary to popular belief, this isn't a reason to pack up your bags and go home. Actually, there are things we can get out of tensors that we can't get out of matrices. We just have to be careful about what types of problems we try to solve. More precisely, in this chapter we will give an algorithm with provable guarantees for low-rank tensor decomposition – that works in natural but restricted settings – as well as some preliminary applications of it to factor analysis.

3.1 The Rotation Problem

Before we study the algorithmic problems surrounding tensors, let's first understand why they're useful. To do this, we'll need to introduce the concept of *factor analysis*, where working with tensors instead of matrices will help us circumvent one of the major stumbling blocks. So, what is factor analysis? It's a basic tool in statistics where the goal is to take many variables and explain them away using a much smaller number of hidden variables, called factors. But it's best to understand it through an example. And why not start with a historical example? It was first used in the pioneering work of Charles Spearman, who had a theory about the nature of intelligence – he believed that there are fundamentally two types of intelligence: *mathematical* and *verbal*. I don't agree, but let's continue anyway.

He devised the following experiment to test out his theory: He measured the performance of one thousand students, each on ten different tests, and

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

arranged his data into a 1000×10 matrix M. He believed that how a student performed on a given test was determined by some hidden variables that had to do with the student and the test. Imagine that each student is described by a two-dimensional vector where the two coordinates give numerical scores quantifying his or her mathematical and verbal intelligence, respectively. Similarly, imagine that each test is also described by a two-dimensional vector, but the coordinates represent the extent to which it tests mathematical and verbal reasoning. Spearman set out to find this set of two-dimensional vectors, one for each student and one for each test, so that how a student performs on a test is given by the inner product between their two respective vectors.

Let's translate the problem into a more convenient language. What we are looking for is a particular factorization

$$M = AB^T$$

where A is size 1000×2 and B is size 10×2 that validates Spearman's theory. The trouble is, even if there is a factorization $M = AB^T$ where the columns of A and the rows of B can be given some *meaningful* interpretation (that would corroborate Spearman's theory) how can we find it? There can be many other factorizations of M that have the same inner dimension but are not the factors we are looking for. To make this concrete, suppose that O is a 2×2 orthogonal matrix. Then we can write

$$M = AB^T = (AO)(O^T B^T)$$

and we can just as easily find the factorization $M = \hat{A}\hat{B}^T$ where $\hat{A} = AO$ and $\hat{B} = BO$ instead. So even if there is a meaningful factorization that would explain our data, there is no guarantee that we find it, and in general what we find might be an arbitrary inner rotation of it that itself is difficult to interpret. This is called the *rotation problem*. This is the stumbling block that we alluded to earlier, which we encounter if we use matrix techniques to perform factor analysis.

What went wrong here is that low-rank matrix decompositions are not unique. Let's elaborate on what exactly we mean by unique in this context. Suppose we are given a matrix M and are promised that it has some meaningful low-rank decomposition

$$M = \sum_{i=1}^{r} a^{(i)} (b^{(i)})^T.$$

Our goal is to recover the factors $a^{(i)}$ and $b^{(i)}$. The trouble is that we could compute the singular value decomposition $M = U\Sigma V^T$ and find another low-rank decomposition

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

$$M = \sum_{i=1}^{r} \sigma_i u^{(i)} (v^{(i)})^T.$$

These are potentially two very different sets of factors that just happen to recreate the same matrix. In fact, the vectors $u^{(i)}$ are necessarily orthonormal, because they came from the singular value decomposition, even though there is a priori no reason to think that the true factors $a^{(i)}$ that we are looking for are orthonormal too. So now we can qualitatively answer the question we posed at the outset. Why are we interested in tensors? It's because they solve the rotation problem and their decomposition is unique under much weaker conditions than their matrix decomposition counterparts.

3.2 A Primer on Tensors

A tensor might sound mysterious, but it's just a collection of numbers. Let's start with the case we'll spend most of our time on. A third-order tensor T has three dimensions, sometimes called *rows*, *columns*, and *tubes*. If the size of T is $n_1 \times n_2 \times n_3$, then the standard notation is that $T_{i,j,k}$ refers to the number in row i, column j, and tube k in T. Now, a matrix is just a second-order tensor, because it's a collection of numbers indexed by two indices. And of course you can consider tensors of any order you'd like.

We can think about tensors many different ways, and all of these viewpoints will be useful at different points in this chapter. Perhaps the simplest way to think of an order-three tensor T is as nothing more than a collection of n_3 matrices, each of size $n_1 \times n_2$, that are stacked on top of each other. Before we go any further, we should define the notion of the rank of a tensor. This will allow us to explore when a tensor is not just a collection of matrices, as well as when and how these matrices are interrelated.

Definition 3.2.1 *A rank-one, third-order tensor T is the tensor product of three vectors u, v, and w, and its entries are*

$$T_{i,j,k} = u_i v_j w_k.$$

Thus if the dimensions of u, v, and w are n_1, n_2, and n_3, respectively, T is of size $n_1 \times n_2 \times n_3$. Moreover, we will often use the following shorthand:

$$T = u \otimes v \otimes w$$

We can now define the rank of a tensor:

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Definition 3.2.2 *The rank of a third-order tensor T is the smallest integer r so that we can write*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}.$$

Recall, the rank of a matrix M is the smallest integer r so that M can be written as the sum of r rank-one matrices. The beauty of the rank of a matrix is how many equivalent definitions it admits. What we have above is the natural generalization of one of the many definitions of the rank of a matrix to tensors. The decomposition above is often called a CANDECOMP/PARAFAC decomposition.

Now that we have the definition of rank in hand, let's understand how a low-rank tensor is not *just* an arbitrary collection of low-rank matrices. Let $T_{\cdot,\cdot,k}$ denote the $n_1 \times n_2$ matrix corresponding to the kth slice through the tensor.

Claim 3.2.3 *Consider a rank-r tensor*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}.$$

Then for all $1 \leq k \leq n_3$,

$$colspan(T_{\cdot,\cdot,k}) \subseteq span(\{u^{(i)}\}_i)$$

and moreover,

$$rowspan(T_{\cdot,\cdot,k}) \subseteq span(\{v^{(i)}\}_i).$$

We leave the proof as an exercise for the reader. Actually, this claim tells us why not every stacking of low-rank matrices yields a low-rank tensor. True, if we take a low-rank tensor and look at its n_3 different slices, we get matrices of dimension $n_1 \times n_2$ with rank at most r. But we know more than that. Each of their column spaces is contained in the span of the vectors $u^{(i)}$. Similarly, their row spaces are contained in the span of the vectors $v^{(i)}$.

Intuitively, the rotation problem comes from the fact that a matrix is just one *view* of the vectors $\{u^{(i)}\}_i$ and $\{v^{(i)}\}_i$. But a tensor gives us multiple views through each of its slices, which helps us resolve the indeterminacy. If this doesn't quite make sense yet, that's all right. Come back to it once you understand Jennrich's algorithm and think about it again.

The Trouble with Tensors

Before we proceed, it will be important to dispel any myths you might have that working with tensors will be a straightforward generalization of working

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

with matrices. So, what is so subtle about working with tensors? For starters, what makes linear algebra so elegant and appealing is how something like the rank of a matrix M admits a number of equivalent definitions. When we defined the rank of a tensor, we were careful to say that what we were doing was taking *one* of the definitions of the rank of a matrix and writing down the natural generalization to tensors. But what if we took a different definition for the rank of a matrix and generalized it in the natural way? Would we get the same notion of rank for a tensor? Usually not!

Let's try it out. Instead of defining the rank of a matrix M as the smallest number of rank-one matrices we need to add up to get M, we could define the rank through the dimension of its column/row space. This next claim just says that we'd get the same notion of rank.

Claim 3.2.4 *The rank of a matrix M is equal to the dimension of its column/row space. More precisely,*

$$rank(M) = dim(colspan(M)) = dim(rowspan(M)).$$

Does this relation hold for tensors? Not even close! As a simple example, let's set $n_1 = k^2$, $n_2 = k$, and $n_3 = k$. Then, if we take the n_1 columns of T to be the columns of a $k^2 \times k^2$ identity matrix, we know that the $n_2 n_3$ columns of T are all linearly independent and have dimension k^2. But the $n_1 n_3$ rows of T have dimension at most k because they live in a k-dimensional space. So for tensors, the dimension of the span of the rows is not necessarily equal to the dimension of the span of the columns/tubes.

Things are only going to get worse from here. There are some nasty subtleties about the rank of a tensor. First, the field is important. Let's suppose T is real-valued. We defined the rank as the smallest value of r so that we can write T as the sum of r rank-one tensors. But should we allow these tensors to have complex values, or only real values? Actually this *can* change the rank, as the following example illustrates.

Consider the following $2 \times 2 \times 2$ tensor:

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

where the first 2×2 matrix is the first slice through the tensor and the second 2×2 matrix is the second slice. It is not hard to show that $\text{rank}_{\mathbb{R}}(T) \geq 3$. But it is easy to check that

$$T = \frac{1}{2}\left(\begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} + \begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix}\right).$$

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

So even though T is real-valued, it can be written as the sum of *fewer* rank-one tensors if we are allowed to use complex numbers. This issue never arises for matrices. If M is real-valued and there is a way to write it as the sum of r rank-one matrices with (possibly) complex-valued entries, there is always a way to write it as the sum of at most r rank-one matrices, all of whose entries are real. This seems like a happy accident, now that we are faced with objects whose rank is field-dependent.

Another worrisome issue is that there are tensors of rank three that can be arbitrarily well-approximated by tensors of rank two. This leads us to the definition of border rank:

Definition 3.2.5 *The* border rank *of a tensor T is the minimum r such that for any $\epsilon > 0$ there is a rank-r tensor that is entrywise ϵ-close to T.*

For matrices, the rank and border rank are the same! If we fix a matrix M with rank r, then there is a finite limit (depending on M) to how well we can approximate it by a rank $r' < r$ matrix. One can deduce this from the optimality of the truncated singular value decomposition for low-rank approximation. But for tensors, the rank and border rank can indeed be different, as our final example illustrates.

Consider the following $2 \times 2 \times 2$ tensor:

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

It is not hard to show that $\text{rank}_{\mathbb{R}}(T) \geq 3$. Yet it admits an arbitrarily good rank-two approximation using the following scheme. Let

$$S_n = \begin{bmatrix} n & 1 \\ 1 & \frac{1}{n} \end{bmatrix}; \begin{bmatrix} 1 & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n^2} \end{bmatrix} \text{ and } R_n = \begin{bmatrix} n & 0 \\ 0 & 0 \end{bmatrix}; \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Both S_n and R_n are rank one, and so $S_n - R_n$ has rank at most two. But notice that $S_n - R_n$ is entrywise $1/n$-close to T, and as we increase n we get an arbitrarily good approximation to T. So even though T has rank three, its border rank is at most two. You can see this example takes advantage of larger and larger cancellations. It also shows that the magnitude of the entries of the best low-rank approximation cannot be bounded as a function of the magnitude of the entries in T.

A useful property of matrices is that the best rank k approximation to M can be obtained directly from its best rank $k+1$ approximation. More precisely, suppose that $B^{(k)}$ and $B^{(k+1)}$ are, respectively, the best rank k and rank $k+1$ approximations to M in terms of, say, Frobenius norm. Then we can obtain $B^{(k)}$ as the best rank k approximation to $B^{(k+1)}$. However, for tensors, the best rank

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

k and rank $k+1$ approximations to T need not share *any* common rank-one terms at all. The best rank k approximation to a tensor is unwieldy. You have to worry about its field. You cannot bound the magnitude of its entries in terms of the input. And it changes in complex ways as you vary k.

To me, the most serious issue at the root of all of this is computational complexity. Of course the rank of a tensor is not equal to the dimension of its column space. The former is *NP*-hard (by a result of Hastad [85]) and the latter is easy to compute. You have to be careful with tensors. In fact, computational complexity is such a pervasive issue, with so many problems that are easy to compute on matrices turning out to be *NP*-hard on tensors, that the title of a well-known paper of Hillar and Lim [86] sums it up: "Most Tensor Problems Are Hard."

To back this up, Hillar and Lim [86] proved that a laundry list of other problems, such as finding the best low-rank approximation, computing the spectral norm, and deciding whether a tensor is nonnegative definite, are *NP*-hard too. If this section is a bit pessimistic for you, keep in mind that all I'm trying to do is set the stage so you'll be as excited as you should be — that there actually is something we can do with tensors!

3.3 Jennrich's Algorithm

In this section, we will introduce an algorithm for computing a minimum rank decomposition that works in a natural but restricted setting. This algorithm is called Jennrich's algorithm. Interestingly, it has been rediscovered numerous times (for reasons that we will speculate on later), and to the best of our knowledge the first place that it appeared was in a working paper of Harshman [84], where the author credits it to Dr. Robert Jennrich.

In what follows, we will assume we are given a tensor T, which we will assume has the following form:

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

We will refer to the factors $u^{(i)}$, $v^{(i)}$, and $w^{(i)}$ as the *hidden* factors to emphasize that we do not know them but want to find them. We should be careful here. What do we mean by find them? There are some ambiguities that we can never hope to resolve. We can only hope to recover the factors up to an arbitrary reordering (of the sum) and up to certain rescalings that leave the rank-one tensors themselves unchanged. This motivates the following definition, which takes into account these issues:

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Definition 3.3.1 *We say that two sets of factors*

$$\left\{(u^{(i)}, v^{(i)}, w^{(i)})\right\}_{i=1}^{r} \text{ and } \left\{(\hat{u}^{(i)}, \hat{v}^{(i)}, \hat{w}^{(i)})\right\}_{i=1}^{r}$$

are equivalent if there is a permutation $\pi : [r] \to [r]$ *such that for all* i

$$u^{(i)} \otimes v^{(i)} \otimes w^{(i)} = \hat{u}^{(\pi(i))} \otimes \hat{v}^{(\pi(i))} \otimes \hat{w}^{(\pi(i))}.$$

The important point is that two sets of factors that are equivalent produce two decompositions

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)} = \sum_{i=1}^{r} \hat{u}^{(i)} \otimes \hat{v}^{(i)} \otimes \hat{w}^{(i)}$$

that have the same set of rank-one tensors in their sums.

The main question in this section is: Given T, can we efficiently find a set of factors that are equivalent to the hidden factors? We will state and prove a version of Jennrich's algorithm that is more general, following the approach of Leurgans, Ross, and Abel [103].

Theorem 3.3.2 *[84], [103] Suppose we are given a tensor of the form*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

where the following conditions are met:

(1) the vectors $\{u^{(i)}\}_i$ *are linearly independent,*
(2) the vectors $\{v^{(i)}\}_i$ *are linearly independent, and*
(3) every pair of vectors in $\{w^{(i)}\}_i$ *is linearly independent.*

Then there is an efficient algorithm to find a decomposition

$$T = \sum_{i=1}^{r} \hat{u}^{(i)} \otimes \hat{v}^{(i)} \otimes \hat{w}^{(i)}$$

and moreover, the factors $(u^{(i)}, v^{(i)}, w^{(i)})$ *and* $(\hat{u}^{(i)}, \hat{v}^{(i)}, \hat{w}^{(i)})$ *are equivalent.*

The original result of Jennrich [84] was stated as a *uniqueness* theorem, that under the conditions on the factors $u^{(i)}$, $v^{(i)}$, and $w^{(i)}$ above, *any* decomposition of T into at most r rank-one tensors must use an equivalent set of factors. It just so happened that the way that Jennrich proved this uniqueness theorem was by giving an algorithm that finds the decomposition, although in the paper it was never stated that way. Intriguingly, this seems to be a major contributor to why the result was forgotten. Much of the subsequent literature cited a

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

stronger uniqueness theorem of Kruskal, whose proof is nonconstructive, and seemed to forget that the weaker uniqueness theorem of Jennrich comes along with an algorithm. Let this be a word of warning: If you not only prove some mathematical fact but your argument readily yields an algorithm, then say so!

Jennrich's Algorithm [84]

Input: tensor $T \in \mathbb{R}^{m \times n \times p}$ satisfying the conditions in Theorem 3.3.2
Output: factors $\{u_i\}_i$, $\{v_i\}_i$, and $\{w_i\}_i$

Choose $a, b \in \mathbb{S}^{p-1}$ uniformly at random; set

$$T^{(a)} = \sum_{i=1}^{p} a_i T_{\cdot,\cdot,i} \text{ and } T^{(b)} = \sum_{i=1}^{p} b_i T_{\cdot,\cdot,i}$$

Compute the eigendecomposition of $T^{(a)}(T^{(b)})^+$ and $((T^{(a)})^+ T^{(b)})^T$

Let U and V be the eigenvectors corresponding to nonzero eigenvalues

Pair up $u^{(i)}$ and $v^{(i)}$ iff their eigenvalues are reciprocals

Solve for $w^{(i)}$ in $T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$

End

Recall that $T_{\cdot,\cdot,i}$ denotes the ith matrix slice through T. Thus $T^{(a)}$ is just the weighted sum of matrix slices through T, each weighted by a_i.

The first step in the analysis is to express $T^{(a)}$ and $T^{(b)}$ in terms of the hidden factors. Let U and V be size $m \times r$ and $n \times r$ matrices, respectively, whose columns are $u^{(i)}$ and $v^{(i)}$. Let $D^{(a)}$ and $D^{(b)}$ be $r \times r$ diagonal matrices whose entries are $\langle w^{(i)}, a \rangle$ and $\langle w^{(i)}, b \rangle$, respectively. Then

Lemma 3.3.3 *$T^{(a)} = UD^{(a)}V^T$ and $T^{(b)} = UD^{(b)}V^T$*

Proof: Since the operation of computing $T^{(a)}$ from T is linear, we can apply it to each of the rank-one tensors in the low-rank decomposition of T. It is easy to see that if we are given the rank-one tensor $u \otimes v \otimes w$, then the effect of taking the weighted sum of matrix slices, where the ith slice is weighted by a_i, is that we obtain the matrix $\langle w, a \rangle u \otimes v$.

Thus by linearity we have

$$T^{(a)} = \sum_{i=1}^{r} \langle w^{(i)}, a \rangle u^{(i)} \otimes v^{(i)}$$

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

which yields the first part of the lemma. The second part follows analogously with a replaced by b. ■

It turns out that we can now recover the columns of U and the columns of V through a generalized eigendecomposition. Let's do a thought experiment. If we are given a matrix M of the form $M = UDU^{-1}$ where the entries along the diagonal matrix D are distinct and nonzero, the columns of U will be eigenvectors, except that they are not necessarily unit vectors. Since the entries of D are distinct, the eigendecomposition of M is unique, and this means we can recover the columns of U (up to rescaling) as the eigenvectors of M.

Now, if we are instead given two matrices of the form $A = UD^{(a)}V^T$ and $B = UD^{(b)}V^T$, then if the entries of $D^{(a)}(D^{(b)})^{-1}$ are distinct and nonzero, we can recover the columns of U and V (again up to rescaling) through an eigendecomposition of

$$AB^{-1} = UD^{(a)}(D^{(b)})^{-1}U^{-1} \text{ and } (A^{-1}B)^T = VD^{(b)}(D^{(a)})^{-1}V^{-1}$$

respectively. It turns out that instead of actually forming the matrices above, we could instead look for all the vectors v that satisfy $Av = \lambda_v Bv$, which is called a generalized eigendecomposition. In any case, this is the main idea behind the following lemma, although we need to take some care, since in our setting the matrices U an V are not necessarily square, let alone invertible matrices.

Lemma 3.3.4 *Almost surely, the columns of U and V are the unique eigenvectors corresponding to nonzero eigenvalues of* $T^{(a)}(T^{(b)})^+$ *and* $((T^{(a)})^+T^{(b)})^T$, *respectively. Moreover, the eigenvalue corresponding to* $u^{(i)}$ *is the reciprocal of the eigenvalue corresponding to* $v^{(i)}$.

Proof: We can use the formula for $T^{(a)}$ and $T^{(b)}$ in Lemma 3.3.3 to compute

$$T^{(a)}(T^{(b)})^+ = UD^{(a)}(D^{(b)})^+U^+$$

The entries of $D^{(a)}(D^{(b)})^+$ are $\langle w^{(i)}, a\rangle / \langle w^{(i)}, b\rangle$. Then, because every pair of vectors in $\{w^{(i)}\}_i$ is linearly independent, we have that almost surely over the choice of a and b, the entries along the diagonal of $D^{(a)}(D^{(b)})^+$ will all be nonzero and distinct.

Now, returning to the formula above for $T^{(a)}(T^{(b)})^+$, we see that it is an eigendecomposition and, moreover, that the nonzero eigenvalues are distinct. Thus the columns of U are the unique eigenvectors of $T^{(a)}(T^{(b)})^+$ with nonzero eigenvalue, and the eigenvalue corresponding to $u^{(i)}$ is $\langle w^{(i)}, a\rangle / \langle w^{(i)}, b\rangle$. An identical argument shows that the columns of V are the unique eigenvectors of

$$((T^{(a)})^+T^{(b)})^T = VD^{(b)}(D^{(a)})^+V^+$$

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

with nonzero eigenvalue. And by inspection, we have that the eigenvalue corresponding to $v^{(i)}$ is $\langle w^{(i)}, b\rangle / \langle w^{(i)}, a\rangle$, which completes the proof of the lemma. ■

Now, to complete the proof of the theorem, notice that we have only recovered the columns of U and the columns of V up to rescaling – i.e., for each column, we recovered the corresponding unit vector. We will push this rescaling factor in with the missing factors $w^{(i)}$. Thus the linear system in the last step of the algorithm clearly has a solution, and what remains is to prove that this is its only solution.

Lemma 3.3.5 *The matrices* $\left\{u^{(i)}(v^{(i)})^T\right\}_{i=1}^{r}$ *are linearly independent.*

Proof: Suppose (for the sake of contradiction) that there is a collection of coefficients that are not all zero where

$$\sum_{i=1}^{r} \alpha_i u^{(i)}(v^{(i)})^T = 0.$$

Suppose (without loss of generality) that $\alpha_1 \neq 0$. Because by assumption the vectors $\{v^{(i)}\}_i$ are linearly independent, we have that there is a vector a that satisfies that $\langle v^{(1)}, a\rangle \neq 0$ but is orthogonal to all other $v^{(i)}$s. Now, if we right multiply the above identity by a, we get

$$\alpha_1 \langle v^{(1)}, a\rangle u^{(1)} = 0$$

which is a contradiction, because the left-hand side is nonzero. ■

This immediately implies that the linear system over the $w^{(i)}$'s has a unique solution. We can write the linear system as an $mn \times r$ matrix, each of whose columns represents a matrix $u^{(i)}(v^{(i)})^T$ but in vector form, times an unknown $r \times p$ matrix whose columns represent the vectors $w^{(i)}$. The product of these two matrices is constrained to be equal to an $mn \times p$ matrix whose columns represent each of the p matrix slices through the tensor T, but again in vector form. This completes the proof of Theorem 3.3.2.

If you want a nice open question, note that the conditions in Jennrich's algorithm can only ever hold if $r \leq \min(n_1, n_2)$, because we need that the vectors $\{u^{(i)}\}_i$ and $\{v^{(i)}\}_i$ are linearly independent. This is called the undercomplete case, because the rank is bounded by the largest dimension of the tensor. When r is larger than either n_1, n_2, or n_3, we know that the decomposition of T is generically unique. But are there algorithms for decomposing generic overcomplete third-order tensors? This question is open even when $r = 1.1 \max(n_1, n_2, n_3)$.

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

3.4 Perturbation Bounds

This section is good medicine. What we have so far is an algorithm (Jennrich's algorithm) that decomposes a third-order tensor T under some natural conditions on the factors, but under the assumption that we know T *exactly*. In our applications, this just won't be enough. We'll need to handle noise. The aim of this section is to answer the question: If we are given $\widetilde{T} = T + E$ instead (you can think of E as representing sampling noise), how well can we approximate the hidden factors?

Our algorithm won't change. We will still use Jennrich's algorithm. Rather, what we want to do in this section is track how the errors propagate. We want to give quantitative bounds on how well we approximate the hidden factors, and the bounds we give will depend on E and properties of T. The main step in Jennrich's algorithm is to compute an eigendecomposition. Naturally, this is where we will spend most of our time – in understanding when eigendecompositions are stable. From this, we will easily be able to see when and why Jennrich's algorithm works in the presence of noise.

Prerequisites for Perturbation Bounds

Now let's be more precise. The main question we're interested in is the following:

Question 5 *If $M = UDU^{-1}$ is diagonalizable and we are given $\widetilde{M} = M + E$, how well can we estimate U?*

The natural thing to do is to compute a matrix that diagonalizes $\widetilde{M}$ – i.e., $\widetilde{U}$, where $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$ – and quantify how good $\widetilde{U}$ is as an estimate for U. But before we dive right in, it's good to do a thought experiment.

There are some cases where it just is not possible to say that U and $\widetilde{U}$ are close. For example, if there are two eigenvalues of M that are very close to each other, then the perturbation E could in principle collapse two eigenvectors into a single two-dimensional eigenspace, and we would never be able to estimate the columns of U. What this means is that our perturbation bounds will have to depend on the minimum separation between any pair of eigenvalues of M.

Just like this, there is one more thought experiment we can do, which tells us another property of M that must make its way into our perturbation bounds. But before we get there, let's understand the issue in a simpler setup. This takes us to an important notion from numerical linear algebra.

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Definition 3.4.1 *The* condition number *of a matrix U is defined as*

$$\kappa(U) = \frac{\sigma_{\max}(U)}{\sigma_{\min}(U)}$$

where $\sigma_{\max}(U)$ and $\sigma_{\min}(U)$ are the maximum and minimum singular values of U, respectively.

The condition number captures how errors amplify when solving systems of linear equations. Let's be more precise: Consider the problem of solving for x in $Mx = b$. Suppose we are given M exactly, but we only know an estimate $\widetilde{b} = b + e$ of b. How well can we approximate x?

Question 6 *If we obtain a solution $\widetilde{x}$ that satisfies $M\,\widetilde{x} = \widetilde{b}$, how close is $\widetilde{x}$ to x?*

We have $\widetilde{x} = M^{-1}\widetilde{b} = x + M^{-1}e = x + M^{-1}(\widetilde{b} - b)$. So

$$\|x - \widetilde{x}\| \leq \frac{1}{\sigma_{\min}(M)}\|b - \widetilde{b}\|.$$

Since $Mx = b$, we also have $\|b\| \leq \sigma_{\max}(M)\|x\|$. It follows that

$$\frac{\|x - \widetilde{x}\|}{\|x\|} \leq \frac{\sigma_{\max}(M)}{\sigma_{\min}(M)}\frac{\|b - \widetilde{b}\|}{\|b\|} = \kappa(M)\frac{\|b - \widetilde{b}\|}{\|b\|}.$$

The term $\|b-\widetilde{b}\|/\|b\|$ is often called the *relative error* and is a popular distance to measure closeness in numerical linear algebra. What the discussion above tells us is that the condition number controls the relative error when solving a linear system.

Now let's tie this back in to our earlier discussion. It turns out that our perturbation bounds for eigendecompositions will also have to depend on the condition number of U. Intuitively, this is because, given U and U^{-1}, finding the eigenvalues of M is like solving a linear system that depends on U and U^{-1}. This can be made more precise, but we won't do so here.

Gershgorin's Disk Theorem and Distinct Eigenvalues

Now that we understand what sorts of properties of M should make their way into our perturbation bounds, we can move on to actually proving them. The first question we need to answer is: Is $\widetilde{M}$ diagonalizable? Our approach will be to show that if M has distinct eigenvalues and E is small enough, then $\widetilde{M}$ also has distinct eigenvalues. The main tool in our proof will be a useful fact from numerical linear algebra called Gershgorin's disk theorem:

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Theorem 3.4.2 *The eigenvalues of an $n \times n$ matrix M are all contained in the following union of disks in the complex plane:*

$$\bigcup_{i=1}^{n} D(M_{ii}, R_i)$$

where $D(a,b) := \{x \mid \|x - a\| \leq b\} \subseteq \mathbb{C}$ and $R_i = \sum_{j \neq i} |M_{ij}|$.

It is useful to think about this theorem in a special case. If $M = I + E$ where I is the identity matrix and E is a perturbation that has only small entries, Gershgorin's disk theorem is what tells us the intuitively obvious fact that the eigenvalues of M are all close to one. The radii in the theorem give quantitative bounds on how close to one they are. Now for the proof:

Proof: Let (x, λ) be an eigenvector-eigenvalue pair (note that this is valid even when M is not diagonalizable). Let i denote the coordinate of x with the maximum absolute value. Then $Mx = \lambda x$ gives $\sum_j M_{ij} x_j = \lambda x_i$. So $\sum_{j \neq i} M_{ij} x_j = \lambda x_i - M_{ii} x_i$. We conclude:

$$|\lambda - M_{ii}| = \left| \sum_{j \neq i} M_{ij} \frac{x_j}{x_i} \right| \leq \sum_{j \neq i} |M_{ij}| = R_i.$$

Thus $\lambda \in D(M_{ii}, R_i)$. ■

Now we can return to the task of showing that $\widetilde{M}$ is diagonalizable. The idea is straightforward and comes from digesting a single expression. Consider

$$U^{-1} \widetilde{M} U = U^{-1}(M + E)U = D + U^{-1} E U.$$

What does this expression tell us? The right-hand side is a perturbation of a diagonal matrix, so we can use Gershgorin's disk theorem to say that its eigenvalues are close to those of D. Now, because left multiplying by U^{-1} and right multiplying by U is a similarity transformation, this in turn tells us about $\widetilde{M}$'s eigenvalues.

Let's put this plan into action and apply Gershgorin's disk theorem to understand the eigenvalues of $\widetilde{D} = D + U^{-1} E U$. First, we can bound the magnitude of the entries of $\widetilde{E} = U^{-1} E U$ as follows. Let $\|A\|_\infty$ denote the matrix max norm, which is the largest absolute value of any entry in A.

Lemma 3.4.3 $\|\widetilde{E}\|_\infty \leq \kappa(U) \|E\|$

Proof: For any i and j, we can regard $\widetilde{E}_{i,j}$ as the quadratic form of the ith row of U^{-1} and the jth column of U on E. Now, the jth column of U has Euclidean

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

norm at most $\sigma_{max}(U)$, and similarly the ith row of U^{-1} has Euclidean norm at most $\sigma_{max}(U^{-1}) = 1/\sigma_{min}(U)$. Together, this yields the desired bound. ■

Now let's prove that, under the appropriate conditions, the eigenvalues of $\widetilde{M}$ are distinct. Let $R = \max_i \sum_j |\widetilde{E}_{i,j}|$ and let $\delta = \min_{i \neq j} |D_{i,i} - D_{j,j}|$ be the minimum separation of the eigenvalues of D.

Lemma 3.4.4 *If $R < \delta/2$, then the eigenvalues of $\widetilde{M}$ are distinct.*

Proof: First we use Gershgorin's disk theorem to conclude that the eigenvalues of $\widetilde{D}$ are contained in disjoint disks, one for each row. There's a minor technicality, that Gershgorin's disk theorem works with a radius that is the sum of the absolute values of the entries in a row, except for the diagonal entry. But we leave it as an exercise to check that the calculation still goes through.

Actually, we are not done yet.[1] Even if Gershgorin's disk theorem implies that there are disjoint disks (one for each row) that contain the eigenvalues of $\widetilde{D}$, how do we know that no disk contains more than one eigenvalue and that no disk contains no eigenvalues? It turns out that the eigenvalues of a matrix are a continuous function of the entries, so as we trace out a path

$$\gamma(t) = (1 - t)D + t(\widetilde{D})$$

from D to $\widetilde{D}$ as t goes from zero to one, the disks in Gershgorin's disk theorem are always disjoint and no eigenvalue can jump from one disk to another. Thus, at $\widetilde{D}$ we know that there really is exactly one eigenvalue in each disk, and since the disks are disjoint, we have that the eigenvalues of $\widetilde{D}$ are distinct as desired. Of course the eigenvalues of $\widetilde{D}$ and $\widetilde{M}$ are the same, because they are related by a similarity transformation. ■

Comparing the Eigendecompositions

We now know that $\widetilde{M}$ has distinct eigenvalues, so we are finally allowed to write $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$, because $\widetilde{M}$ is diagonalizable. Let's turn to our final step. There is a natural correspondence between eigenvalues of M and eigenvalues of $\widetilde{M}$, because what the proof in the previous subsection told us was that there is a collection of disjoint disks that contains exactly one eigenvalue of M and exactly one eigenvalue of $\widetilde{M}$. So let's permute the eigenvectors of $\widetilde{M}$ to make our life notationally easier. In fact, why not make it easier still. Let's assume (without loss of generality) that all the eigenvectors are unit vectors.

[1] Thanks to Santosh Vempala for pointing out this gap in an earlier version of this book. See also [79].

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Now suppose we are given $(\widetilde{u}_i, \widetilde{\lambda}_i)$ and (u_i, λ_i), which are corresponding eigenvector-eigenvalue pairs for $\widetilde{M}$ and M, respectively. Let $\sum_j c_j u_j = \widetilde{u}_i$. We know that there is a choice of c_j's that makes this expression hold, because the u_j's are a basis. What we want to show is that in this expression, c_j for all $j \neq i$ is small. This would imply that u_i and $\widetilde{u}_i$ are close.

Lemma 3.4.5 *For any $j \neq i$, we have*

$$|c_j| \leq \frac{\|E\|}{\sigma_{min}(U)(\delta - R)}.$$

Proof: We'll get this by manipulating the expression $\sum_j c_j u_j = \widetilde{u}_i$. First, multiplying both sides of the equation by $\widetilde{M}$ and using the fact that $\{u_i\}_i$ are eigenvectors of M and $\{\widetilde{u}_i\}_i$ are eigenvectors of $\widetilde{M}$, we get

$$\sum_j c_j \lambda_j u_j + E\widetilde{u}_i = \widetilde{\lambda}_i \widetilde{u}_i$$

which, rearranging terms, yields the expression $\sum_j c_j(\lambda_j - \widetilde{\lambda}_i)u_j = -E\widetilde{u}_i$.

Now what we want to do is pick out just one of the coefficients on the left-hand side and use the right-hand side to bound it. To do this, let w_j^T be the j^{th} row of U^{-1}, and left multiplying both sides of the expression above by this vector, we obtain

$$c_j(\lambda_j - \widetilde{\lambda}_i) = -w_j^T E \widetilde{u}_i.$$

Now let's bound the terms in this expression. First, for any $i \neq j$, we have $|\lambda_j - \widetilde{\lambda}_i| \geq |\lambda_j - \lambda_i| - R \geq \delta - R$ using Gershgorin's disk theorem. Second, $\widetilde{u}_i$ is a unit vector by assumption and $\|w_j\| \leq 1/\sigma_{min}(U)$. Using these bounds and rearranging terms now proves the lemma. ■

The three lemmas we have proven can be combined to give quantitative bounds on how close U is to $\widetilde{U}$, which was our goal at the outset.

Theorem 3.4.6 *Let M be an $n \times n$ matrix with eigendecomposition $M = UDU^{-1}$. Let $\widetilde{M} = M + E$. Finally, let*

$$\delta = \min_{i \neq j} |D_{i,i} - D_{j,j}|$$

i.e., the minimum separation of eigenvalues of M.

(1) If $\kappa(U)\|E\|n < \frac{\delta}{2}$, then $\widetilde{M}$ is diagonalizable.

(2) Moreover, if $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$, then there is a permutation $\pi : [n] \to [n]$ such that for all i

$$\|u_i - \widetilde{u}_{\pi(i)}\| \leq \frac{2\|E\|n}{\sigma_{min}(U)(\delta - \kappa(U)\|E\|n)}$$

where $\{u_i\}_i$ are the columns of U and $\{\widetilde{u}_i\}_i$ are the columns of $\widetilde{U}$.

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

Proof: The first part of the theorem follows by combining Lemma 3.4.3 and Lemma 3.4.4. For the second part of the theorem, let's fix i and let P be the projection onto the orthogonal complement of u_i. Then, using elementary geometry and the fact that the eigenvectors are all unit vectors, we have

$$\|u_i - \widetilde{u}_{\pi(i)}\| \leq 2\|P\widetilde{u}_{\pi(i)}\|.$$

Moreover, we can bound the right-hand side as

$$\|P\widetilde{u}_{\pi(i)}\| = \Big\| \sum_{j \neq i} c_j P u_j \Big\| \leq \sum_{j \neq i} |c_j|.$$

Lemma 3.4.5 supplies the bounds on the coefficients c_j, which completes the proof of the theorem. ■

You were warned early on that the bound would be messy! It is also by no means optimized. But what you should instead take away is the qualitative corollary that we were after: If $\|E\| \leq \text{poly}(1/n, \sigma_{min}(U), 1/\sigma_{max}(U), \delta)$ (i.e., if the sampling noise is small enough compared to the dimensions of the matrix, the condition number of U, and the minimum separation), then U and $\widetilde{U}$ are close.

Back to Tensor Decompositions

Now let's return to Jennrich's algorithm. We've stated enough messy bounds for my taste. So let's cheat from here on out and hide messy bounds using the following notation: Let

$$A \xrightarrow{E \to 0} B$$

signify that as E goes to zero, A converges to B at an inverse polynomial rate. We're going to use this notation as a placeholder. Every time you see it, you should think that you could do the algebra to figure out how close A is to B in terms of E and various other factors we'll collect along the way.

With this notation in hand, what we want to do is *qualitatively* track how the error propagates in Jennrich's algorithm. If we let $\widetilde{T} = T + E$, then $\widetilde{T} \xrightarrow{E \to 0} T$ and $\widetilde{T}^{(a)} \xrightarrow{E \to 0} T^{(a)}$ where $\widetilde{T}^{(a)} = \sum_i a_i \widetilde{T}_{\cdot,\cdot,i}$. We leave it as an exercise for the reader to check that there are natural conditions where

$$(\widetilde{T}^{(b)})^+ \xrightarrow{E \to 0} (T^{(b)})^+.$$

As a hint, this convergence depends on the smallest singular value of $T^{(b)}$. Or, to put it another way, if E is not small compared to the smallest singular value of $T^{(b)}$, then in general we cannot say that $(T^{(b)})^+$ and $(\widetilde{T}^{(b)})^+$ are close.

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

In any case, combining these facts, we have that

$$\widetilde{T}^{(a)}(\widetilde{T}^{(b)})^{+} \xrightarrow{E\to 0} T^{(a)}(T^{(b)})^{+}.$$

Now we are in good shape. The eigenvectors of the right-hand side are the columns of U. Let the columns of $\widetilde{U}$ be the eigenvectors of the left-hand side. Since the left-hand side is converging to the right-hand side at an inverse polynomial rate, we can invoke our perturbation bounds on eigendecompositions (Theorem 3.4.6) to conclude that their eigenvectors are also converging at an inverse polynomial rate. In particular, $\widetilde{U} \xrightarrow{E\to 0} U$ where we have abused notation, because the convergence above is only after we have applied the appropriate permutation to the columns of $\widetilde{U}$. Similarly, we have $\widetilde{V} \xrightarrow{E\to 0} V$.

Finally, we compute $\widetilde{W}$ by solving a linear system in $\widetilde{U}$ and $\widetilde{V}$. It can be shown that $\widetilde{W} \xrightarrow{E\to 0} W$ using the fact that $\widetilde{U}$ and $\widetilde{V}$ are close to well-conditioned matrices U and V, which means that the linear system we get from taking the tensor product of the ith column in $\widetilde{U}$ with the ith column in $\widetilde{V}$ is also well-conditioned.

These are the full, gory details of how you can prove that Jennrich's algorithm behaves well in the presence of noise. If we make our life easy and in what follows analyze our learning algorithms in the no-noise case ($E = 0$), we can always appeal to various perturbation bounds for eigendecompositions and track through how all the errors propagate to bound how close the factors we find are to the true hidden factors. This is what I meant by good medicine. You don't need to think about these perturbation bounds every time you use tensor decompositions, but you should know that they exist, because they really are what justifies using tensor decompositions for learning problems where there is always sampling noise.

3.5 Exercises

Problem 3-1:

(a) Suppose we want to solve the linear system $Ax = b$ (where $A \in \mathbb{R}^{n\times n}$ is square and invertible) but we are only given access to a noisy vector $\tilde{b}$ satisfying

$$\frac{\|b - \tilde{b}\|}{\|b\|} \leq \varepsilon$$

and a noisy matrix $\tilde{A}$ satisfying $\|A - \tilde{A}\| \leq \delta$ (in operator norm). Let $\tilde{x}$ be the solution to $\tilde{A}\tilde{x} = \tilde{b}$. Show that

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press

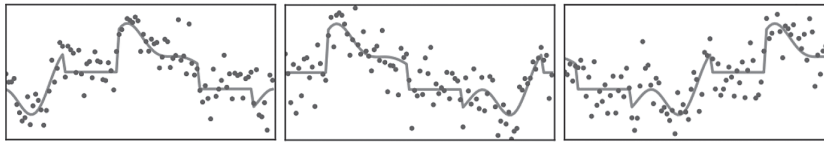

Figure 3.1: Three shifted copies of the true signal x are shown in gray. Noisy samples y_i are shown in red. (Figure credit: [23].)

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\varepsilon\, \sigma_{\max}(A) + \delta}{\sigma_{\min}(A) - \delta}$$

provided $\delta < \sigma_{\min}(A)$.

(b) Now suppose we know A exactly, but A may be badly conditioned or even singular. We want to show that it may still be possible to recover a specific coordinate x_j of x. Let $\tilde{x}$ be any solution to $A\tilde{x} = \tilde{b}$ and let a_i denote column i of A. Show that

$$|x_j - \tilde{x}_j| \leq \frac{\|b - \tilde{b}\|}{C_j}$$

where C_j is the norm of the projection of a_j onto the orthogonal complement of span $(\{a_i\}_{i \neq j})$.

Problem 3-2: In the *multireference alignment* problem, we observe many noisy copies of the same unknown signal $x \in \mathbb{R}^d$, but each copy has been circularly shifted by a random offset (Figure 3.1).
Formally, for $i = 1, 2, \ldots, n$ we observe

$$y_i = R_{\ell_i} x + \xi_i$$

where the ℓ_i are drawn uniformly and independently from $\{0, 1, \ldots, d-1\}$; R_ℓ is the operator that circularly shifts a vector by ℓ indices; $\xi_i \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$ with $\{\xi_i\}_i$ is independent; and $\sigma > 0$ is a known constant. Think of d, x, and σ as fixed while $n \to \infty$. The goal is to recover x (or a circular shift of x).

(a) Consider the tensor $T(x) = \frac{1}{d}\sum_{\ell=0}^{d-1}(R_\ell x) \otimes (R_\ell x) \otimes (R_\ell x)$. Show how to use the samples y_i to estimate T (with error tending to zero as $n \to \infty$). Take extra care with the entries that have repeated indices (e.g., T_{aab}, T_{aaa}).

(b) Given $T(x)$, prove that Jennrich's algorithm can be used to recover x (up to circular shift). Assume that x is *generic* in the following sense: Let $x' \in \mathbb{R}^d$ be arbitrary and let x be obtained from x' by adding a small perturbation $\delta \sim \mathcal{N}(0, \epsilon)$ to the first entry. *Hint:* Form a matrix with rows $\{R_\ell x\}_{0 \leq \ell < d}$, arranged so that the diagonal entries are all x_1.

https://doi.org/10.1017/9781316882177.004 Published online by Cambridge University Press