

Deep Learning
Prof. Mithesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

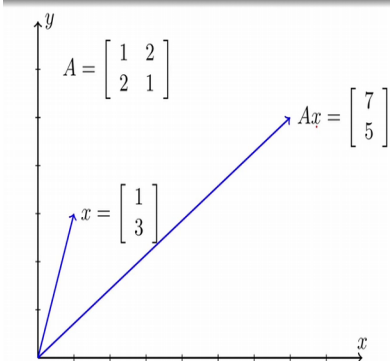
Lecture - 06
Eigen Values, Eigen Vectors, Eigen Value Decomposition, Principal Component Analysis, Singular Value Decomposition

So, this lecture actually is a bit of a digression, and it is supposed to cover some of the basics that we need for various sections of the course. So, it is very important that you understand some concepts for linear algebra specifically eigenvalues, eigenvectors and in particular. Today we will do principal component analysis, and the reason that I do it is there is an very neat relation of PCA and to autoencoders. An autoencoder is something that we'll cover in the course, it is a part of any deep neural network course.

And singular value decomposition is something that we use when we learn word vectors. The word vector is again something very important, I can just do the non SVD version of it where I just talk about what word is, but that will not give you the same probably not the same interpretation as if you start from SVD and then reach word vectors, right. So, that is why I am covering these basics.

So, how many of you know eigenvalues and eigenvectors? Very embarrassing question how many of you absolutely hate eigenvalues and eigenvectors. So, let us see if we can change that today ok, I mean on the positive side right.

(Refer Slide Time: 01:24)



$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

$Ax = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$

- What happens when a matrix hits a vector?
- The vector gets transformed into a new vector (it strays from its path)
- The vector may also get scaled (elongated or shortened) in the process.

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

So, what happens when a matrix hits a vector? So, most of you a lot of people that I talk to right actually think that eigenvectors are the villains of linear algebra, it is very hard to understand them and so on. But today I am going to make a case for they are not the villains they are actually the superheroes of linear algebra. So, that is what the lecture is about ok. So, what happens when a matrix hits a vector?

Student: Transforms it.

Transforms it right; so, actually what happens is that it strays from its path. So, this is the original (Refer Time: 01:58) this is the original vector x and now once I multiply it by A ; that means, if I do the transformation Ax then I get a new vector. And two things happen right, one is the direction changes which is obvious, and in many cases the scale also changes; that means, the vector might get elongated its magnitude would increase or it would decrease right.

So, if you really think about it actually right. So, matrices are the real villains of linear algebra right, and we just look at this vector was minding its own business going along its own direction a matrix comes and hits it and completely changes its world right, I mean; it just throws it off path increases a dimension or slows it down or whatever it. So, that is they are the bad guys now for every villain what do you have a superhero right. So, what is a superhero corresponding to a matrix? What does a superhero do? Know that is

a very linear algebra. I am talking about comic books that this is very linear algebraic answer he stands up to the villain right ok.

(Refer Slide Time: 02:54)

$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$Ax = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

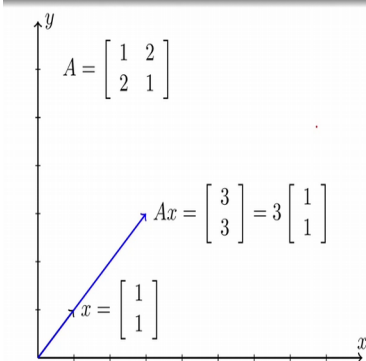
- For a given square matrix A , there exist special vectors which refuse to stray from their path.
- These vectors are called eigenvectors.
- More formally,
 $Ax = \lambda x$ [direction remains the same]
- The vector will only get scaled but will not change its direction.

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

And that is exactly what Eigen vectors do it right, they refused to change their part they tell the matrix ok. You can hit me as many times as you want probably you can increase my you could probably slow me down a bit or push me ahead or something, but I am not going to stray off from your path right. So, that is what eigenvalue eigenvectors do.

So here is a matrix, which is a villain and here is an eigenvector which is our hero and now when this matrix hits this eigenvector it refuses to stray from it is part right. It says I will move forward I will move back whatever, but I will not change my direction ok. I will just stay honest to what I am and these vectors are called the eigenvectors. I am more formally you can write it as Ax is equal to λx right so; that means, the direction remains the same only the scale changes it will either get slowed down or it will get boosted up right. So, the magnitude would change, but the direction remains the same ok.

(Refer Slide Time: 03:50)



$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$Ax = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

- So what is so special about eigenvectors?
- Why are they always in the limelight?
- It turns out that several properties of matrices can be analyzed based on their eigenvalues (for example, see spectral graph theory)
- We will now see two cases where eigenvalues/vectors will help us in this course

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

Now, what is so special about eigenvectors? Like why are they always in the lime light? I know the any course that you do invariably touch eigenvectors or eigenvalues at some point in that course right, where the beat machine learning image processing whatever you do you always speech everything that you do, you will always have eigenvectors and eigenvalues, why is I? So, well it is turns out that several properties of matrices can actually be explained away by looking at their eigenvalues right. So, if I look at a matrix I would probably not be able to comment much on it, but if you tell me something about the eigenvalues.

I can see a lot of things about of it and there is an entire field on this way this entire spectral graph theory which looks at properties of Laplacian matrices and come in something on the properties of the graph and so on right. And that is just an example which we do not care about, but what we care about in this course there are a few things that we care about with respect to eigenvalues and eigenvector. And that is what I am going to focus on right. So, that is what this lecture is going to be out. And I will take 2 specific cases which are very important for us to understand certain concepts later on. So, I will start with the first one.

(Refer Slide Time: 04:50)

Chinese Mexican

k_1
 k_2

$$v_{(0)} = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

$$v_{(1)} = \begin{bmatrix} pk_1 + (1-q)k_2 \\ (1-p)k_1 + qk_2 \end{bmatrix}$$

$$= \begin{bmatrix} p & 1-q \\ 1-p & q \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

$$v_{(1)} = Mv_{(0)}$$

$$v_{(2)} = Mv_{(1)}$$

$$= M^2v_{(0)}$$

In general, $v_{(n)} = M^n v_{(0)}$

- Let us assume that on day 0, k_1 students eat Chinese food, and k_2 students eat Mexican food. (Of course, no one eats in the mess!)
- On each subsequent day i , a fraction p of the students who ate Chinese food on day $(i-1)$, continue to eat Chinese food on day i , and $(1-p)$ shift to Mexican food.
- Similarly a fraction q of students who ate Mexican food on day $(i-1)$ continue to eat Mexican food on day i , and $(1-q)$ shift to Chinese food.
- The number of customers in the two restaurants is thus given by the following series:

$$v_{(0)}, Mv_{(0)}, M^2v_{(0)}, M^3v_{(0)}, \dots$$

6/71

Prof. Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 6

And I will start with a very simple example to motivate this problem. And eventually will lead to a result which will help us understand a very important concept in deep neural network training which is exploding and vanishing, vanishing reboots. We will not touch that concept today, but we will use these ideas when we are looking at that later on.

So, let us take this example of 2 restaurants. So, there is a Chinese restaurant and a Mexican restaurant. And on day one k_1 students eat in the Chinese restaurant and k_2 students eat in the Mexican restaurant ok. So, this is what my situation is on day 0, k_1 for Chinese and k_2 for Mexican ok. Now what happens as is obvious people get bored or they have different want to try out different things. So, on day two or other each subsequent day what happens is that, a fraction p of the students who ate Chinese today will offer max Mexican, on day on the next day and a fraction q of the students who ate ma Mexican today are going to offer Chinese.

So, you get this situation right. So, I started with k_1 k_2 . So, what I am saying is on day one that is the next day only a fraction p of the k_1 students will remain for Chinese and a fraction $1 - q$ would be transferred from Mexican to Chinese ok. And similarly only a fraction q of the students would again stick to the Mexican food and a fraction $1 - p$ into k_1 would shift from Chinese to Mexican is this setup clear ok. Can you write this as a matrix operation it would be a matrix multiplied by a vector right can you tell me the vector.

Student: (Refer Time: 06:29).

k_1 k_2 k_1 k_2 and the matrix is in all this ok, this is what it is. And I am saying that this happens on each subsequent day, it is every day now this keeps happening. So, on day 1 I started with say 180 and now day 2 it change to something again day 3 it will change something by the same fraction.

Now, let me call this as matrix M and this is of course, v_0 right by definition as we decided now what would happen on day 2 what would v_2 be M applied to v_1 right and which would be M square applied to v_0 . I am just substituting the value of v_1 which is M into v_0 in general on the n th day what would happen M raised to n into v_0 ok. So, you see that the number of customers in the 2 restaurants is given by this series you had v_0 then M into v_0 then M square v_0 and so on up to M raised to n v_n ok. You see how the number of customer is changing.

Now, and this is how I represent it as a state transition diagram right. So, I had certain numbers on day 1 and it changed with the trans with the probability p they will stay back with a probability $1 - p$ they will move to the next or the different restaurant and so on right.

(Refer Slide Time: 07:32)

```
graph LR; k1((k1)) -- p --> k1; k1 -- 1-p --> k2((k2)); k2 -- q --> k2; k2 -- 1-q --> k1;
```

- This is a problem for the two restaurant owners.
- The number of patrons is changing constantly.
- Or is it? Will the system eventually reach a steady state? (i.e. will the number of customers in the two restaurants become constant over time?)
- Turns out they will!
- Let's see how?

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

And now this though a very toyish example can you relate it to many things in real life or many things that you will take in decision making rate that you are. So, even if you are

playing a game for example, and even if you are playing Atari games or something, you are in a certain state based on some action that will take will move to a different state and so on right. So, these things happen in various real world applications right there is a certain state for example, even in stock market prediction, you are at a certain value of fish stock it might change to a different value right and these values you could just say them as high low or neutral that I am not going into the actual numbers.

Today the stock value is high it does it possibility that it will transition to something low and so on right. So, these kind of straight transition diagrams occur in various real world examples. Now this is a problem for the two restaurant owners right, why is this a problem for the two restaurant owners? They do not know how much food to make, but every day the number of customers is changing right, but is the number of customers actually changing. Will the system eventually reach a steady state? Will it is it obvious that it will reach a steady state or maybe it will not even reaches steady, but the way I describe it I do not see why it should reach a steady state right you have some people here they go there come back go there and so on.

The only thing which I have assumed is that the transition matrix which was the matrix M is constant across all the time steps right. So, every day it is at the same priorities by which things are changed right. So, what is your guess if I were to ask you to take a guess ok. Let us see how many of you think and it is there is no correct answer here at this point. So, just tell me how many of you think it will reach a steady state? How many of you think it will keep changing and why is the sum never equal to 1 ok. So, fine so it turns out that they will right and let us see how.

(Refer Slide Time: 09:40)

Definition
Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues of an $n \times n$ matrix A . λ_1 is called the dominant eigen value of A if

$$|\lambda_1| \geq |\lambda_i| \quad i = 2, \dots, n$$

Definition
A matrix M is called a stochastic matrix if all the entries are positive and the sum of the elements in each column is equal to 1.
(Note that the matrix in our example is a stochastic matrix)

Theorem
The largest (dominant) eigenvalue of a stochastic matrix is 1.
[See proof here](#)

Theorem
If A is a $n \times n$ square matrix with a dominant eigenvalue, then the sequence of vectors given by $Av_0, A^2v_0, \dots, A^nv_0, \dots$ approaches a multiple of the dominant eigenvector of A .
(the theorem is slightly misstated here for ease of explanation)

8/71

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

So, we will define some things and some of these are just definitions some of them have accompanying proofs, which I am not going to do here you can the proofs have been linked from the slides. So, you can take a look at them if you are interested right.

So, suppose there is a matrix A n cross n matrix which has eigenvalues are $\lambda_1, \lambda_2, \dots, \lambda_n$. Now what this definition is saying is that, assume that there is one eigenvalue which is greater there is no assumption actually the eigenvalue which is greater than all the other eigenvalues is called the dominant eigenvalue. And when I am looking at a dominant eigenvalue I am only concerned with the magnitude not the sign ok. So, it could be that an eigenvalue is minus 10 and all the other eigenvalues are 1 2 3 4 5. So, the dominant eigenvalue would be minus 10 right and I will just take it as step is that clear the definition of a dominant eigenvalue ok.

Now, how many of you know what is the stochastic matrix? So, matrix M is called a stochastic matrix, if all the entries are positive and the sum of the elements in each column is equal to 1. So now, this definition is again slightly misstated. So, there is a row stochastic matrix the column stochastic matrix and also doubly stochastic matrix right. So, what I am talking about here is a column stochastic matrix like our matrix have you seen such a stochastic matrix any time in your life in the last 5 minutes the M matrix right. So, the M matrix is a stochastic matrix because the sum of the columns was 1 right,

you had $p + 1 - p - q + 1 - q$ or was it some of the rows was 1 rows was 1 is it the columns fine.

So, this is a stochastic matrix just a definition. Now I combine these two definitions which is, dominant eigenvalue and stochastic matrix and give you a theorem right. So, the largest dominant or the dominant eigenvalue of a stochastic matrix is equal to 1 ok. So, to prove this, what do I have to prove? So, I need to prove two things one that 1 is an eigenvalue of this matrix of any stochastic matrix and second all the other eigenvalues are less than 1. So, that is exactly what this proof does here you can take a look at it and just to give you a heads up. So, last year I use to do this that please see the proof go back and look at the proof people never look at the proofs.

So, I used to ask them in the quiz where I should be sure that people not going to answer right. So, please when I say go back and look at the proof do that ok. So, and lastly if A is an $n \times n$ square matrix and you have this series $A^0 v, A^1 v, A^2 v, \dots, A^n v$, then this series will converge to the dominant eigenvector of A . What does a statement mean? Let us not get into the proof right what does it actually mean ok. So, let us start with very basic stuff it what is the series actually? What is each element in this series it is a vector, it is a vector everyone gets that every element in the series is a vector?

Now what do I mean that a series of vectors converges to the dominant Eigen vector, what is convergence mean? If I keep finding the next element next element next element of this series and I keep doing this as long as I can. I will reach a value n right where n is the n th element in the series which will just be a multiple of the dominant Eigen vector is that clear? You not seem to be clear everyone gets that ok.

So, what do you mean by if you take a series of numbers and if I say that the series converges to 0, what does that mean? If you keep finding the next element in the series, you will hit a point n where you find the n th element of the series and it will be 0 (Refer Time: 13:20) that ok. So, we will just I will leave it at that for now. Now so stochastic matrix dominant eigenvalues the connection between 2 and the convergence theorem for a series of vectors which is $A^0 v, A^1 v, A^2 v, \dots$ and so on ok.

(Refer Slide Time: 30:36)

- Let e_d be the dominant eigenvector of M and $\lambda_d = 1$ the corresponding dominant eigenvalue
- Given the previous definitions and theorems, what can you say about the sequence $Mv_{(0)}, M^2v_{(0)}, M^3v_{(0)}, \dots$?
- There exists an n such that

$$v_{(n)} = M^n v_{(0)} = ke_d \text{ (some multiple of } e_d)$$
- Now what happens at time step $(n + 1)$?

$$v_{(n+1)} = Mv_{(n)} = M(ke_d) = k(Me_d) = k(\lambda_d e_d) = ke_d$$
- The population in the two restaurants becomes constant after time step n .
[See Proof Here](#)

```

graph LR
    k1((k1)) -- p --> k1
    k1 -- 1-p --> k2((k2))
    k2 -- q --> k2
    k2 -- 1-q --> k1
  
```

Prof. Mitesh M. Khapra | CS7015 (Deep Learning) : Lecture 6

Now, let e_d be the dominant Eigen vector of M where M is a dash matrix in our case it is a stochastic matrix. So, what with the corresponding dominant eigenvalue be.

Student: 1.

1 so given the previous definitions and theorems, what can you say about the sequence? It converges to a dash of e_d .

Student: (Refer Time: 13:59).

A multiple of e_d right. So, there exists an n such that the a length n th element of the series which is given by this is going to be equal to some multiple of the dominant eigenvector no, no; k is some multiple no this is not related to eigenvalues yet just wait for the next statement, then you will see the difference that this is not the do eigenvalue yet ok.

Now, my question is what happens from here onwards, what would be the next element in the series. How many of you say some k dash into e_d ? What is the other pause I do not have the other option what is the other option.

Student: k into e_d .

k into e_d how many of you say k into e_d ? A large number of ok so, you see that now just notice the eigenvalue will come up right. So, at step n plus 1 you would have M into v_n

which is M into k into e_d and this quantity is actually 1. So, the theorem says it will converge to some multiple of k and now if it is a stochastic matrix, what will happen after that time step? It will just remain the same vector.

So, what would happen to the number of customers in the two restaurants it will remain the same right you get that fine. Now, this was all for, what kind of matrices? Stochastic matrices square stochastic matrices ok.

(Refer Slide Time: 15:15)

- Now instead of a stochastic matrix let us consider any square matrix A
- Let p be the time step at which the sequence x_0, Ax_0, A^2x_0, \dots approaches a multiple of e_d (the dominant eigenvector of A)



$$A^p x_0 = k e_d$$

$$A^{p+1} x_0 = A(A^p x_0) = k A e_d = k \lambda_d e_d$$

$$A^{p+2} x_0 = A(A^{p+1} x_0) = k \lambda_d A e_d = k \lambda_d^2 e_d$$

$$A^{p+n} x_0 = k (\lambda_d)^n e_d$$

- In general, if λ is the dominant eigenvalue of a matrix A , what would happen to the sequence x_0, Ax_0, A^2x_0, \dots if
 - $|\lambda| > 1$ (will explode)
 - $|\lambda| < 1$ (will vanish)
 - $|\lambda| = 1$ (will reach a steady state)

NPTEL Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

But we generally care about any square matrix. In fact, we should care about any matrix not discriminate, but any square matrix will do for now. So, for a square matrix let p be the time step at which this series approaches a multiple of the dominant eigenvector.

The theorem was for any square matrix, remember it was not for stochastic square matrices. We just use this value that for a stochastic square matrix the dominant eigenvalue is 1, which it need which leads to that neat result that the num then the number of customers just becomes constant right ok. But for any square matrix, I could write it as this that there exist some step p at which the element of the peath element of the series would just be a multiple of the dominant eigenvector ok.

Now, what would happen at step p plus 1? Is this fine ok, what about step p plus 2, and in general at p plus k or p plus n everyone gets this ok. So now, can you tell me what does

this knowing this dominant Eigen value tell us about this series, when will it stabilize actually?

Student: (Refer Time: 16:25).

When lambda is equal to 1 that is the case we already saw if the dominant Eigen value is greater than 1, what would happen?

Student: (Refer Time: 16:33).

Series will explode the series will explode and if it is less than one what would happen the series will vanish ok. So, this is an important result that we will use when we are discussing exploding and managing gradients.

So, we will see that in the case of something one as a recurrent neural networks, you end up with something of this sort and then I will make some comments on that right. So, that is why we will be using this will come probably 6 7 or maybe more lectures down the line ok, but we will be using it at this point. So, the main result from here is that if the dominant eigenvalue, this should be λ_d is greater than 1. Then it will explode less than 1 it will vanish and equal to 1 it will stabilize ok, is that fine ok. So, that is one result one important property of eigenvalues and eigenvectors that we'll be needing at a later point in the course.