Syllabus

0.1 Admin Info

Class Name: Linear Algebra 2 (E)
Class Time: Th 19:20-20:55
Class Location: 4-4106
Instructor: Yilong Yang (YMSC)
Email: yy26@mail.tsinghua.edu.cn
Office: Jinchunyuan West 137
Office Hours: Th 5PM-6PM
TA: Lovy Singhal
TA Office: TBD
TA Office Hour: TBD
Discussion Session Time: TBD
Discussion Session Location: TBD
Class Wechat Group: TBD

0.2 Prerequisite

You should have mastered the following materials and skills:

1. Linear Combination and Linear Dependency.
2. Gausssian Elimination and LU decomposition.
3. Row and Column Operations to find Reduced Row Echelon Forms, to solve Linear Systems, and to find Determinants.
4. Matrix Inversion and Multiplication.
5. The Fundamental Theorem of Linear Algebra (Rank-nullity theorem and orthogonality between the four fundamental subspaces of a matrix.)
6. Gram-Schmidt Orthogonalization and QR-Decomposition.
7. Projections and Orthogonal Projections.
8. Change of Basis and orthogonal change of basis.
9. Eigenvalues and Eigenvectors.
10. Criteria for Diagonalizability.
11. Spectral Theorem for Real Symmetric Matrices.
12. Singular Value Decomposition.

In the stuff listed above, I want to specifically stress that, even though we do NOT need singular value decomposition in this class, it is probably HIGHLY IMPORTANT that you know it. It has tons of applications and will likely show up in your future.

If you do not know about singular value decompositions, you can read Gilbert Strang's introduction to linear algebra, chapter seven. (There are also accompanying online videos from MIT open course if you like.)

0.3 Content of Class

Textbook:

[LN] My lecture notes. This year's will be updated as our class go along, but feel free to check last year's notes.

[OLN] My old lecture notes, written up in 2019 Spring.

[OV] Online videos from 2020 spring, made during the COVID-19 pandemic.

Optional Textbook:

[GS] Gilbert Strang, Introduction to Linear Algebra 5th edition. The linear algebra textbook used in MIT. (University bookstore) This is NOT the main textbook, but we shall use some sections of it.

[ST] Sergei Treil, Linear Algebra Done Wrong. The linear algebra textbook used in Brown University for honor linear algebra class, and the one I used when I was a freshman. (Author made it free online.) We shall use some sections from it.

[SA] Sheldon Axler, Linear Algebra Done Right. Great linear algebra textbooks for math majors. A bit too hardcore sometimes.

[NH] Nicholas J Higham, Functions of Matrices: Theory and Computation. The first two chapters are all we need.

[BW] Ray M. Bowen and C. C. Wang, Introduction to Vectors and Tensors. Good for the tensor portion of the class.

Content Structure:

1. Complex Matrices (GS Ch 9)
2. Jordan Normal Form (ST Ch 9, SA Ch 8)
3. Matrix Analysis (NH ch 1)
4. Dual and Tensor (LADW Ch 8 and Lecture notes)
5. (Optional) ??? if we have time.

0.4 Grading

30% Homework, 30% Midterm, 30% Final, and 10% Project.

Homework: The homeworks should usually due weekly. Tries to write in english, but we do not really test your english ability, and it is totally fine if you let slide some Chinese if you are really struggling to express yourself.

All answers must be supplimented with proofs unless specifically told not to. Proofs need not to be rigiorous, but it is your job to make your reasoning clear to the grader. The grader should not be banging his/her head trying to decipher your logic. You are welcome to come to me or the TA for grading disputes.

As far as deadlines go, I'm usually easygoing, but I reserve the right to refuse any late submission.

Midterm: You take it home, you do it for two weeks, and you hand them back. Sort of like a glorified homework, but you must hand them in on time. The problems will of course be very hard. You will likely lose some hair.

Final: Open book final on our last class. (The university do not assign standard final exam times for "special" classes such as ours.) The time is tentatively 7PM-10PM. It will be significantly easier and more standardized than the Midterm.

Project: TBD. Mostly this would be some self-learning projects.

Collaboration:

I think stress is detrimental to all learning endeavor, and competition is meaningless in a classroom setting since all of us have the same goal, to learn. As a general principle, I encourage collaboration of all sorts.

Ideally, I hope that you look at the problems as soon as I put them up, and think independently at first. You do NOT need to do them right away. Look them first, think a little bit, and maybe sleep on them for a day or two. As you can see from the grading policy, I tried my best to minimize your stress, so you can take your time and think them through. Some problems are DESIGNED so that you might need a few days to solve. After a day or two, if the answer still eludes you, feel free to ask your classmates for collaboration.

I encourage collaborations on homeworks, projects and even the takehome midterm. However, you must obey the following rule:

1. You MUST each hand in your own work individially in your own words.
2. You MUST understand everything you wrote. (Say you copied your friend's WRONG answer without thinking, and that will most likely be in violation of this rule.)
3. You need to write down the names of your collaborator.
4. Failure to comply rule 2 and rule 3 will be treated as plagiarism.
5. Collaboration with people not in this class (such as a math grad student) is not forbidden but not recommended. If you choose to, then write down their names as well.

0.5 Classroom Policy

1. You are allowed to sleep, eat, drink during class as long as no other classmate objects to it. (Unless a school official come to observe. Then please be on your best behavior wink wink wink.)
2. We do not record attendance, but coming to class is obviously highly recommended, especially since I do extra stuff all the time and they will be tested.
3. You may speak or interrupt me without raising your hand at all time during class. If my writing, speaking or explaining confuses you somehow, it is very admirable of you to speak up about it.
4. Respect your classmates. Which means turn your phone to vibrate in class; admire them rather than judge them when your classmates ask questions in class; and when asked to collaborate, assume that they are competent and want to learn, and explain and discuss patiently with them. Do not insult your classmate by just throwing your answers to them, as if they are not worthy of your time, or as if they are hopelessly stupid to figure things out.

0.6 Class Schedule

Contents

Part I

Complex Matrix Theory

Chapter 1

Complex Matrices

1.1 What is a complex linear combination?

We are entering into the second portion of your linear algebra education, and we are going to see more complex matrices. A complex matrix is, in a very nominal sense, a matrix with possibly complex entries, say $\begin{bmatrix} 1+\mathrm{i} & -\mathrm{i} \\ 2-\mathrm{i} & 3 \end{bmatrix}$. But this should NOT be satisfactory for you, because what does it even mean?

Let us do a little review first.

Recall that a matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$ is representing a linear map. In particular, it represents some process that respect linear combinations. As a quick example, say we are playing a version of the famous board game "settlers of catan". If you want to build a road, you would need to spend one wood and one brick. If you want to build a ship, you would need to spend one wood and one wool. So if you want to build $\begin{bmatrix} x \text{ roads} \\ y \text{ ships} \end{bmatrix}$, then you would need $A \begin{bmatrix} x \text{ roads} \\ y \text{ ships} \end{bmatrix} = \begin{bmatrix} x+y \text{ woods} \\ x \text{ bricks} \\ y \text{ wools} \end{bmatrix}$. So A is the evaluation process that tells you how much your required building would cost. This process is LINEAR, because the total cost of "a linear combination of buildings" is the linear combination of the cost of each type of building. It RESPECTS the linear combination in the sense that $A(s\boldsymbol{v} + t\boldsymbol{w}) = s(A\boldsymbol{v}) + t(A\boldsymbol{w})$.

If you forget all about our class last quarter, at least I hope you would remember these. A vector is representing a linear combination, and a matrix is representing a linear map, which is a map that preserves linear combinations. (Personally I think this perspectives on linear combinations and linear maps is WHY we learn linear algebra in college. No other stuff is not important.)

Now, under this view, the idea of a complex matrix like $\begin{bmatrix} 1+\mathrm{i} & -\mathrm{i} \\ 2-\mathrm{i} & 3 \end{bmatrix}$ is very disturbing. This seems to be about COMPLEX linear combinations, in contrast the the real linear combinations that we are used to. It is very easy to imagine the likes of "two apples and three bananas", but what is the meaning of an imaginary apple? So before we move on, we need a little extra perspective on complex numbers and complex linear combinations.

First of all, why do we even need complex numbers? The answer is obvious: we want a degree n polynomial to have an n-th root. This is straightforward enough. Over the reals, $x^2 + 1 = 0$ has no solution, which is super annoying. For example, without complex numbers, $\begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}$ has NO eigenvector and no eigenvalues, which is annoying. But over complex numbers, it will have distinct eigenvalues $\pm\mathrm{i}$, and in fact it will be diagonalizable. Hooray!

So this establishes the necessity of complex numbers. But where can we go search for this? As you recall in your high school complex number class, to have the complex numbers, all we need is to find the imaginary i, which is a square root of -1. With this square root of minus one, we can then have all complex numbers.

So the meaning of complex numbers ultimately depends on the meaning of the imaginary unit i. What is the meaning of this i?

Example 1.1.1. We are searching for x such that $x^2 = -1$. But broaden our minds a little bit. Can we find a matrix A such that $A^2 = -I$?

Yes we can. Consider the 2×2 real matrices, which are linear transformations on $\mathbb{R}^2$, the plane. On the plane, what is $-I$? That is basically reflecting everything about the origin, i.e., rotation by 180 degree. So what operation A can we find, such that A^2 is rotation by 180 degree? The answer is rotation by 90 degree, easy.

I hope you still remembered how to find this matrix. The answer is (if we rotate counter-clockwise) $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Of course, $-A$ also satisfies $(-A)^2 = -I$, so we in fact have at least two solutions, $\pm A$, just like $x^2 = -1$ has two solutions, $\pm i$. (We in fact have infinitely many solutions to the matrix equations $A^2 = -I$. Can you find a way to describe them all?)

Now is time to witness magic. Lo and behold the wonders of algebra.

$$(2 + 3\mathrm{i})(4 + \mathrm{i}) = 5 + 14\mathrm{i}.$$

$$\begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 5 & -14 \\ 14 & 5 \end{bmatrix}.$$

Why is this even true? Let me explain this by rewriting the second equation, and then I'll leave the thinking to you.

$$\begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ 1 & 4 \end{bmatrix} = (2I + 3A)(4I + A) = 5I + 14A.$$

Let me end this exploration with one question for you to think. Suppose some $n \times n$ matrix A satisfies $A^2 = -I$, then would we have a similar structure? ☺

Example 1.1.2. Bonus foods for your thought. Compute the following two matrix multiplications. What would you get? How are the two following calculations related?

$$\begin{bmatrix} 1 & \mathrm{i} \\ 2\mathrm{i} & 1+\mathrm{i} \end{bmatrix} \begin{bmatrix} \mathrm{i} & 1-\mathrm{i} \\ 2 & \mathrm{i} \end{bmatrix} = ?$$

$$\left[\begin{array}{cc|cc} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ \hline 0 & -2 & 1 & -1 \\ 2 & 0 & 1 & 1 \end{array}\right] \left[\begin{array}{cc|cc} 0 & -1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ \hline 2 & 0 & 0 & -1 \\ 0 & 2 & 1 & 0 \end{array}\right] = ?$$

Suppose some $n \times n$ matrix A satisfies $A^2 = -I$, then can you construct similar coincidences? ☺

Example 1.1.3. We have hinted that whenever $A^2 = -I$, then you can choose i as representing A, and use complex numbers. What are other possible A? Here is an exotic (but useful) example.

Let V be the space of functions of the form $a\sin(x) + b\cos(x)$. Let $A : V \to V$ be the linear map of taking derivatives. Then note that $A^2 = -I$ in this space. ☺

The above serves to point out that the imaginary unit i has very real meanings, and possibly many many meanings, and you should pick your own meaning depending on the application at hand. Luckily for us, most of the time, when people use complex numbers, they are usually interpreting the imaginary i as some sort of rotation, i.e., $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

Under this interpretation, a complex number $a + b\mathrm{i}$ can be interpreted as $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. So a pure real number is like a dilation operation on the plane, while a purely imaginary number is like a rotation operation on the plane. Here is a example copied from the book "One Two Three ... Infinity".

Example 1.1.4. A treasure is buried on an island. To find the treasure, we start at a location with a flag (location Z). We then first walk to a building (location A), say with a total distance of x, then we turn right and walk x. Let us call this location A'.

Next we go back to the flag (location Z). We then first walk to a statue (location B), say with a total distance of y, then we turn left and walk y. Let us call this location B'.

The treasure is at the midpoint between A' and B'.

Now some bad guy came and took away the flag (so Z is unknown). Can you still find the treasure? Yes we can.

Note that $A' - A$ is $A - Z$ rotated clockwise, so $A' - A = -\mathrm{i}(A - Z)$. Similarly, $B' - B$ is $B - Z$ rotated counter-clockwise, so $B' - B = \mathrm{i}(B - Z)$. So the treasure locatio $\frac{1}{2}(A' + B') = \frac{1}{2}(A + B) + \frac{1}{2}\mathrm{i}(B - A)$, and no Z is involved in this. So the flag position does not matter at all. I'll leave the interpretation of the final treasure location to yourself.

This is NOT showing you the power of complex numbers. Rather, this is showing you the power of linear algebra. At the center of the entire calculation is the fact that rotation is linear. The complex numbers such as i are merely names that we slap on the operations such as rotations.

So... linear algebra rules, and complex numbers are just names and labels for convenience. ☺

So, when we are dealing with objects that can be "rotated", it would make sense to talk about i times that object. In this sense, we can do complex-linear combinations. No wonder that quantum mechanics where using complex numbers.

All in all, for a complex vector such as $\boldsymbol{v} = \begin{bmatrix} 1 \\ \mathrm{i} \\ 1-\mathrm{i} \end{bmatrix}$, it is better to think of each coordinate as representing a point in the plane. And if we perform a complex scalar multiplication $(2 + \mathrm{i})\boldsymbol{v}$, think of this as applying a planar operation $\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ to each coordinate of $\boldsymbol{v}$.

Here are some other fun applications of complex numbers.

Example 1.1.5 (Complex romantic relation)**.** Suppose $f' = kf$, then I'm sure you know that the solution is $f(x) = \mathrm{e}^{kx} f(0)$. That is the prerequisite knowledge of this application.

Suppose two person A, B are in a romantic relation.Their love for each other is a function of time, say $A(t)$ and $B(t)$. Now A is a normal person. For normal people, the more you are loved, the more you love back. In particular, $A'(t) = B(t)$. However, B is an unappreciative person. If you love B, then B take you for granted, and treat you as garbage. If, however, you treat B badly, then B would all of a sudden thinks of you as super charming and attractive. In short, B enjoys things that are hard to get, and think little of the things that are easy to get. In Chinese, we say B is a Jian Ren. Anyway, we see that $B'(t) = -A(t)$.

Now, consider the real vector $\boldsymbol{v}(t) = \begin{bmatrix} A(t) \\ B(t) \end{bmatrix} \in \mathbb{R}^2$. Then for the matrix $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, we see that $\boldsymbol{v}' = -J\boldsymbol{v}$. Now, think of $\mathbb{R}^2$ as simply $\mathbb{C}$, and v would be like some complex number, and J is the rotation counter-clockwise by 90 degree, i.e., multiplication by i. And we have $\boldsymbol{v}' = -\mathrm{i}\boldsymbol{v}$. So the solution is $\boldsymbol{v}(t) = \mathrm{e}^{-\mathrm{i}t}\boldsymbol{v}(0) = (\cos(t) - \mathrm{i}\sin(t))\boldsymbol{v}(0)$.

Then the solution should be $(\cos(t)I - \sin(t)J)\begin{bmatrix} A(0) \\ B(0) \end{bmatrix} = \begin{bmatrix} A(0)\cos(t) + B(0)\sin(t) \\ B(0)\cos(t) - A(0)\sin(t) \end{bmatrix}$. This is indeed the collection of all possible solutions of our system. We have solved the differential equation.

Note that the romantic relation of A and B are necessarily periodic. If you are ever trapped in a relationship which is periodic, (i.e., happy for a week, then fight for a week, and repeat), then maybe you should think about this model a bit more. ☺

1.2 Complex Orthogonality

Procedural-wise, complex linear algebra works in the same way as real linear algebra. The Gaussian elimination works the same way. The matrix multiplication formula, the trace formula and the determinant formula are all the same. Nothing new all in all. However, one thing is crucially different: inner product, and by extension, transpose.

For two real vectors $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}2\\-1\end{bmatrix}$, it is very easy to understand that they are orthogonal to each other. We can draw it, or visualize it in our mind, and so on. But for two complex vectors, what does it mean to be orthogonal to each other?

Example 1.2.1. Consider $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ and $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$. What would happen if we perform the "real dot-product" on these two vectors? We would have $1^2+(\mathrm{i})^2 = 1+(-1) = 0$. Huh, this vector is "orthogonal" to itself? How can it be?

It simply cannot be. Quoting Sherlock Holmes, when you have eliminated the impossible, whatever remains, however improbable, must be the truth: we used the wrong "dot product"!

There is a lesson we can learn from this. Blindly apply analogous procedures will usually lead you astray. It is always to guide your scientific exploration with proper intuitions.

What is $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$? Recall that previously, we have talked about the relation between $a+b\mathrm{i}$ and $\begin{bmatrix}a&-b\\b&a\end{bmatrix}$. Using this interpretation, let us think of $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ as $\begin{bmatrix}1&0\\0&1\\0&-1\\1&0\end{bmatrix}$. So instead of one vector, it is in fact two vectors!

So what is orthogonal to $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$? Well, let us consider $\begin{bmatrix}1\\-\mathrm{i}\end{bmatrix}$. Then the two vectors can be thought of as $\begin{bmatrix}1&0\\0&1\\0&-1\\1&0\end{bmatrix}$ and $\begin{bmatrix}1&0\\0&1\\0&1\\-1&0\end{bmatrix}$. Did you see that? ALL FOUR column vectors are mutually orthogonal to each other. So we conclude that $\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ and $\begin{bmatrix}1\\-\mathrm{i}\end{bmatrix}$ are orthogonal to each other.

What does this mean? It means that if n-dimensional complex vectors $\boldsymbol{v},\boldsymbol{w}$ corresponds to $2n\times 2$ real matrices A,B, then we say $\boldsymbol{v}\perp\boldsymbol{w}$ if and only if $A^{\mathrm{T}}B$ has all four entries zero.

Something funny is going on here. Note that, by interpreting i as $\begin{bmatrix}0&-1\\1&0\end{bmatrix}$, we are interpreting $\boldsymbol{v}=\begin{bmatrix}1\\\mathrm{i}\end{bmatrix}$ as $A=\begin{bmatrix}1&0\\0&1\\0&-1\\1&0\end{bmatrix}$. Then $A^{\mathrm{T}}=\begin{bmatrix}1&0&0&1\\0&1&-1&0\end{bmatrix}$, and it does NOT represent $\boldsymbol{v}^{\mathrm{T}}$. Rather, it represents $\overline{\boldsymbol{v}}^{\mathrm{T}}$. Here the line means complex conjugates on each coordinate.

In particular, the fact that $A^{\mathrm{T}}B$ is the 2×2 zero meatrix corresponds to the fact that $\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}$ is the complex number zero. ☺

Definition 1.2.2. *For two complex vectors $\boldsymbol{v},\boldsymbol{w}\in\mathbb{C}^n$, then we define their complex dot product to be $\langle\boldsymbol{v},\boldsymbol{w}\rangle=\overline{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{w}$.*

A generic guideline is that, whenever you take transpose for a real matrix, in the corresponding world of complex matrices, you probably would like to take a transpose conjugate. Think of this as a generalization of the following fact: if $\begin{bmatrix}a&-b\\b&a\end{bmatrix}$ represents $a+b\mathrm{i}$, then its transpose actually represents $a-b\mathrm{i}$. For convenience, we shall use the "star" as a shorthand for conjugate transpose, i.e., we define A^* as $\overline{A}^{\mathrm{T}}$.

For example, we have the following result.

Theorem 1.2.3. *For a complex $m\times n$ matrix A, then* $\operatorname{Ran}(A)$ *and* $\operatorname{Ker}(A^*)$ *are orthogonal complements, and* $\operatorname{Ran}(A^*)$ *and* $\operatorname{Ker}(A)$ *are orthogonal complements. Oh, and* $\operatorname{Ran}(A)$ *and* $\operatorname{Ran}(A^*)$ *and* $\operatorname{Ran}(A^{\mathrm{T}})$ *have the same complex dimension, i.e., the rank of A.*

Familar yes? We have a bunch of similar results here. Note that ultimately, everything here involves an orthogonal structure, which is why conjugate transpose is used throughout. Review or read up about their real conterparts if needed.

1. A complex matrix is ***Hermitian*** if $A = A^*$. In this case, it is diagonalizable with real eigenvalues, and the underlying space has an orthogonal basis made of eigenvectors of A.
2. A complex matrix is ***skew-Hermitian*** if $-A = A^*$. In this case, it is diagonalizable with purely-imaginary eigenvalues, and the underlying space has an orthogonal basis made of eigenvectors of A.
3. A complex matrix is ***unitary*** if $A^{-1} = A^*$. In this case, it is diagonalizable with unit complex eigenvalues (complex numbers with absolute value one), and the underlying space has an orthogonal basis made of eigenvectors of A. Note that in particular, such a map would preserve the complex dot product, i.e., $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \langle A\boldsymbol{v}, A\boldsymbol{w} \rangle$.
4. A complex matrix is ***normal*** if $AA^* = A^*A$. In this case, it is diagonalizable, and the underlying space has an orthogonal basis made of eigenvectors of A.

1.3 Fourier Matrix

Here is a family of matrices that is both super cool, extremely useful in practice, and also illustrates some funny situations mentioned above. It is the famous Fourier matrix.

For any n, let ω be the ***primitive n-th root of unity***, i.e., it is the complex number $\omega = \cos(2\pi/n) + \mathrm{i}\sin(2\pi/n)$. Then as you can check, $1, \omega, ..., \omega^{n-1}$ are all distinct complex numbers, and $\omega^n = 1$. In fact, by thinking of complex numbers as dilations and rotations, it is easy to see that $1, \omega, ..., \omega^{n-1}$ are ALL solutions to the equation $x^n = 1$ over the complex numbers.

We start by looking at the fourier matrix F_n whose (i,j) entry is $\omega^{(i-1)(j-1)}$. For a typical example, we have $F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}$.

As you can see, it appears that $F_n^{\mathrm{T}} = F_n$. However, it is NOT Hermitian. (For example, its diagonal is not real.) In fact, it is the opposite of Hermitian: it is a multiple of a unitary matrix. Feel free to perform $F_4F_4^*$ to verify the case when $n = 4$. In particular, you can also check that $\frac{1}{n}\overline{F_n} = F_n^{-1}$.

The fourier matrix is closely related to the Fourier series and Fourier Transforms. In Calculus we learned that Fourier series is very important. For a periodic function $f(x)$ with period 2π, you can try to decompose it into different frequencies via fourier series, and write it as a linear combination of sines and cosines. Say we have maybe $f(x) = \sum c_k \mathrm{e}^{k\mathrm{i}x}$. Here note that $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$, so $\mathrm{e}^{\mathrm{i}x}$ is just a lazy way to write sine and cosine simultaneously.

Suppose we have a decomposition $f(x) = c_0 + c_1\mathrm{e}^{\mathrm{i}x} + c_2\mathrm{e}^{2\mathrm{i}x} + c_3\mathrm{e}^{3\mathrm{i}x}$. Given c_0, c_1, c_2, c_3, what do we know about the function $f(x)$? Well, if you apply F_4 to the vector $\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$, then you can verify that you have $\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix}$. As you can see, you get four points on the graph of $f(x)$. By using more fourier coefficients, and larger Fourier matrix, you will get more detailed points on your graph for $f(x)$. This is the forward direction.

But consider the backward direction as well. In pratical cases, we usually have the graph of $f(x)$ by some data gathering. How can we work out the Fourier coefficients? Suppose we have $f(x) = c_0 +$

$c_1 e^{ix} + c_2 e^{2ix} + c_3 e^{3ix}$ where the c_i are unknown. How to find the fourier coefficient of $f(x)$? We could evaluate $f(0), f(\pi/2), f(\pi), f(3\pi/2)$ empirically or experimentially, and then compute $F_4^{-1}\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix} = \frac{1}{n}\overline{F_4}\begin{bmatrix} f(0) \\ f(\pi/2) \\ f(\pi) \\ f(3\pi/2) \end{bmatrix}$. As you can see, by evaluating at merely a few points and apply $\frac{1}{n}\overline{F_n}$, we can conveniently obtain the (approximate) Fourier coefficients. The approximation will get better as we use more data points and larger Fourier matrix.

Suppose you want to compute the first 1000 fourier coefficients (say you know the rest are probably noises or measurement errors). In effect, you want to quickly multiply F_{1000} to a known vector. Wow, that is pretty big! How should you do it? By brute fource, this is a 1000 by 1000 matrix, and calculating with it needs millions of calculations. That would take forever. So a better approach is the Fast Fourier Transfour. We start by looking at F_{1024}, reduce it to F_{512}, then reduce it to F_{256}, and so forth, until we reach $F_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. So in 10 steps, we reduce the problem to a much smaller one. In the end, one million calculations will be reduced to merely 5000 calculations. Imagine the gain in speed in signal processing and etc. This is ranked as the top 10 algorithms of the 20-th centry by the IEEE journal Computing in Science and Engineering.

Example 1.3.1. Consider $F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}$. Observe the relation between its first and third column, and between its second and forth column. You can see that the first and third coordinates of corresponding columns are the same, and the second and forth coordinates are negated.

Let us now swap the columns to bring the original first and third column togehter, and the original second and forth column together. Then we have $F_4 P_{23} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & i & -i \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -i & i \end{bmatrix}$. Hey, note that the upper left corner and lower left corner is exactly $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = F_2$! In fact, let $D_2 = diag(1, i)$, we have $F_4 P_{23} = \begin{bmatrix} F_2 & D_2 F_2 \\ F_2 & D_2 F_2 \end{bmatrix} = \begin{bmatrix} I_2 & D_2 \\ I_2 & -D_2 \end{bmatrix}\begin{bmatrix} F_2 & 0 \\ 0 & F_2 \end{bmatrix}$. So step by step, we have extracted F_2 out of F_4! ☺

Theorem 1.3.2 (Fast Fourier Transform)**.** *We have the following decomposition, where $D_n = (1, \omega, ..., \omega^{n-1})$ where $\omega = \cos(\pi/n) + i\sin(\pi/n)$, and P is a matrix permuting all odd columns to the left and all even columns to the right.*

$$F_{2n} = \begin{bmatrix} I_n & D_n \\ I_n & -D_n \end{bmatrix}\begin{bmatrix} F_n & 0 \\ 0 & F_n \end{bmatrix} P.$$

Proof. Do it yourself. Same idea as Example 1.3.1. □

Example 1.3.3. Here's what happen after a recursion. You will have

$$F_{4n} = \begin{bmatrix} I_{2n} & D_{2n} \\ I_{2n} & -D_{2n} \end{bmatrix}\begin{bmatrix} I_n & D_n & 0 & 0 \\ I_n & -D_n & 0 & 0 \\ 0 & 0 & I_n & D_n \\ 0 & 0 & I_n & -D_n \end{bmatrix}\begin{bmatrix} F_n & 0 & 0 & 0 \\ 0 & F_n & 0 & 0 \\ 0 & 0 & F_n & 0 \\ 0 & 0 & 0 & F_n \end{bmatrix} P.$$

Here P is a permutation matrix that put all (1 mod 4) columns to the left, followed by the (3 mod 4) columns, followed by the (2 mod 4) columns, and followed by the (4 mod 4) columns. ☺

Proof. Do it yourself. □

Example 1.3.4. What would happen to F_{3n}? Can you do something similar? I'll leave this to yourself. ☺

Chapter 2

Jordan Canonical Form

2.1 Generalized Eigenstuff

We are moving towards Jordan canonical form. For a square matrix A, sometimes it is diagonalizable. And by doing so, we shall find all the eigenvalues and eigenvectors and so on, so that we can completely understand the behavior of this matrix. But what if we cannot diagonalize a matrix?

Well, first let us strive for a block-diagonalization.

2.1.1 (Review) Block Matrices in $\mathbb{R}^n$ or $\mathbb{C}^n$

We use block matrices a lot, and we know that they can be multiplied like regular matrices and so on. But let us be reminded here about their meaning. Block matrices are NOT just a formality in grouping entries. Each individual block is in fact a linear "submap" in some sense.

Example 2.1.1. Consider a map sending foods to nutrients. Say we have foods: apples, bananas, meat. And we have nutrients: fibers, proteins, suger. Then this map is a matrix A, such that if we have $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ apples, bananas and meat, then we have $A\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ fibers, proteins and suger. Obviously A is a 3 by 3 matrix.

Now consider the block form $A = \left[\begin{array}{cc|c} a & b & c \\ \hline d & e & f \\ g & h & i \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where A_{ij} represent the corresponding blocks.

What does A_{11} do? It sends fruits to the low calory nutrients they contain. What does A_{12} do? It send fruits to the high calory nutrients they contain. What does A_{21} do? It sends meat to the low calory nutrients it contains. What does A_{22} do? It send meat to the high calory nutrients it contains.

$$\begin{tikzcd} \text{fruits} \arrow[r, "A_{11}"] \arrow[dr, "A_{21}"] & \text{low calory} \\ \text{meat} \arrow[r, "A_{22}"] \arrow[ur, "A_{12}"] & \text{high calory} \end{tikzcd}.$$

And what is A? A as a linear map is simply the collection of these four linear maps. ☺

Intuitively, when we have a block matrix, we are grouping input coordinates and output coordinates. The block A_{ij} records how the j-th group of inputing coordinates effect the i-th group of outputing coordinates.

Example 2.1.2. Consider $\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 1 & 2 \\ \hline 0 & 0 & 1 \end{array}\right]$. Note that the lower left block is zero. This means the first two input coodinates does NOT effect the third output coordiante.

Indeed we have $\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 1 & 2 \\ \hline 0 & 0 & 1 \end{array}\right] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x+y+z \\ x+y+2z \\ z \end{bmatrix}$.

$$\begin{tikzcd} \mathbb{R}^2 \arrow[r, "A_{11}"] & \mathbb{R}^2 \\ \mathbb{R} \arrow[ur, "A_{12}"] \arrow[r, "A_{22}"] & \mathbb{R} \end{tikzcd}.$$

This is a ***block upper triangular matrix***.

In particular, block diagonal means each groups of coordinates only effect themselves. In particular, instead of one system, it is more like many separate independent systems, one for each diagonal block. Here is a picture for $\left[\begin{array}{cc|c} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$, which is a ***block diagonal matrix***.

$$\begin{tikzcd} \mathbb{R}^2 \arrow[r, "A_{11}"] & \mathbb{R}^2 \\ \mathbb{R} \arrow[r, "A_{22}"] & \mathbb{R} \end{tikzcd}.$$

As you can see, a block diagonal matrix happens exactly when the two "linear submaps" are independent of each other. ☺

So here is how one can think about block matrices. For example, for the block matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ where A is $m_1 \times n$ and B is $m_2 \times n$, we can think of it as this:

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "A"] \arrow[dr, "B"] & \mathbb{R}^{m_1} \\ & \mathbb{R}^{m_2} \end{tikzcd}.$$

And for the block matrix $\begin{bmatrix} A & B \end{bmatrix}$ where A is $m \times n_1$ and B is $m \times n_2$, we can think of it as this:

$$\begin{tikzcd} \mathbb{R}^{n_1} \arrow[r, "A"] & \mathbb{R}^m \\ \mathbb{R}^{n_2} \arrow[ur, "B"] & \end{tikzcd}.$$

Now, why would the block matrices multiply exactly as regular matrices? Let us reprove this via more diagrams. We have $\begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1B_1 + A_2B_2$ because of this:

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, bend left, "B_1"] \arrow[rr, bend left, "B_2"] & \mathbb{R}^a \arrow[rrr, bend right, "A_1"'] & \oplus & \mathbb{R}^b \arrow[r, bend right, "A_2"'] & \mathbb{R}^m \end{tikzcd}$$

And we have $\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2 \end{bmatrix}$ because of this:

$$\begin{tikzcd}
\mathbb{R}^{n_1} \arrow[dr, "B_1"] & & \mathbb{R}^{m_1} \\
& \mathbb{R}^d \arrow[ur, "A_1"] \arrow[dr, "A_2"] & \\
\mathbb{R}^{n_2} \arrow[ur, "B_2"] & & \mathbb{R}^{m_2}
\end{tikzcd}.$$

Example 2.1.3. Consider a rotation in $\mathbb{R}^3$ around the line $x = y = z$ that sends the positive x-axis to the positive y-axis, and the positive y-axis to the positive z-axis, and the positive z-axis to the positive x-axis. How to find the matrix R of this linear map?

By looking at the standard basis, we obviously have $R = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. If we break down the domain and codomain as a sum of three one-dimensional subspaces, i.e., the coordinate-axes, then we have a diagram:

$$\begin{tikzcd}
\mathbb{R}_x \arrow[dr] & \mathbb{R}_x \\
\mathbb{R}_y \arrow[dr] & \mathbb{R}_y \\
\mathbb{R}_z \arrow[uur] & \mathbb{R}_z
\end{tikzcd}.$$

The arrows here are identity maps. And the arrows NOT DRWAN are zero maps.

Let us try a different decomposition of the domain and the codomain. What if we think of the domain and codomain as the sum of the xy-plane and the z-axis? Then we shall have a block structure $R = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix}$ where R_1 is a 2×2 matrix, and R_2 is 1×2, and R_3 is 2×1, and R_4 is 1×1.

To find R_1, we want to understand the action of R on the xy-plane, ignoring the z-axis. So we want to look at the projection of $R\boldsymbol{e}_1, R\boldsymbol{e}_2$ back to the xy-plane. Since the positive x-axis goes to the positive y-axis, and the positive y-axis goes to the positive z-axis (which is projected to the origin), we see that $R_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. You can work out the others similarly, and you shall have $R = \left[\begin{array}{cc|c} 0 & 0 & 1 \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \end{array}\right]$. ☺

We use mostly $\mathbb{R}^n$ here, but it does not really matter. Replace them all by $\mathbb{C}^n$ if you like.

2.1.2 (Review) Spatial Decompositions and invariant decompositions

We are now going to reformulate everything in the last section in an abstract manner.

Example 2.1.4. Recall that we say V is the direct sum of its subspaces V_1, V_2 if $V_1 \cap V_2 = \{0\}$, and $V_1 + V_2 = V$. We also write $V = V_1 \oplus V_2$, and call this a decomposition of V into subspaces. Now, there are four linear maps involved in this structure.

First of all, we have an inclusion map $\iota_1 : V_1 \to V$ and $\iota_2 : V_2 \to V$. These maps don't change the input at all, but their codomain is larger than the domain. They tell us how the smaller spaces (the domains) is included in the bigger space (the codomain).

Now since $V = V_1 \oplus V_2$, by our knowledge in the last semester, each vector $\boldsymbol{v} \in V$ has a UNIQUE decomposition $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ such that $\boldsymbol{v}_i \in V_i$. So we also have two projection maps $p_1 : V \to V_1$ and $p_2 : V \to V_2$ such that $p_i(\boldsymbol{v}) = \boldsymbol{v}_i$. These are INDEED projection maps. For example, note that for any $\boldsymbol{v}_1 \in V_1$, then $\boldsymbol{v}_1 = \boldsymbol{v}_1 + \boldsymbol{0}$ must be the unique decomposition according to $V = V_1 \oplus V_2$. Therefore $p_1(\boldsymbol{v}_1) = \boldsymbol{v}_1$. In particular, $p_i^2 = p_i$. (This is the defining algebraic property for projections in any mathematical context.) However, these are NOT necessarily orthogonal projections. They could be oblique projections. See last semester's note for oblique projections. (They are only orthogonal projections when $V_1 \perp V_2$. Otherwise they are oblique projections, where p_i preserves V_i and kills V_j for $j \neq i$.)

Now if we have a linear map $L : V \to W$, and decompositions $V = V_1 \oplus V_2$ and $W = W_1 \oplus W_2$. Then there are four possible linear maps induced from these structures. We can restrict the domain of L to V_i and project the codomain to W_j, and obtain $L_{ij} = p_j \circ L \circ \iota_i : V_i \to W_j$. Then we can write $L = \begin{bmatrix} L_{11} & L_{21} \\ L_{12} & L_{22} \end{bmatrix}$. For each $\boldsymbol{v} \in V$, if the unique decomposition according to $V = V_1 \oplus V_2$ is $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$, then let us write it as $\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix}$, and and we do similar things in W. Then we shall see that $\begin{bmatrix} L_{11} & L_{21} \\ L_{12} & L_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} (L\boldsymbol{v})_1 \\ (L\boldsymbol{v})_2 \end{bmatrix}$. ☺

These whole venture is purely philosophical, and you need to feel no pressure to master these abstract computations. My goal is to address the following question: What is the idea behind a block matrix? It means that as we decompose domain and codomain into subspaces, the linear map is decomposed into submaps. The "blocks" are actually "submaps", or restrictions of the original linear map to corresponding subspaces.

Now we go back to our task of block diagonalizing matrices.

Why are diagonal matrices neat? Consider $\begin{bmatrix} d_1 & & \\ & d_2 & \\ & & d_3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} d_1 a_1 \\ d_2 a_2 \\ d_3 a_3 \end{bmatrix}$. As you can see, for a diagonal matrix, treated as a linear map, it acts on each coordinate independently. The i-th coordinate of the output depends only on the i-th coordinate of the input, and vice versa, the i-th coordinate of the input will influence only the i-th coordinate of the output. Coordinates will NOT cross-influence each other, they just each do their own thing during this linear map.

Given a diagonalizable matrix, how would we diagonalize it? We need to find eigenvectors. Each eigenvector is like an invariant direction that the matrix must preserve. Now our matrix acts on each invaraint direction independently, so if we pick a basis made of eigenvectors, then our matrix after a corresponding change of basis will be diagonal.

Now, invariant directions are like one dimensional invariant subspaces. In general, we can define the following:

Definition 2.1.5. *We say a subspace W of a space V is an* ***invariant subspace*** *of the linear transformation $L : V \to V$ if $L(W) \subseteq W$. (We do NOT require them to be equal. The point is such that L can be restricted to a linear transformation on W.)*

We say a decomposition $V = V_1 \oplus V_2$ is an invariant decomposition for the linear transformation $L : V \to V$ if both V_1 and V_2 are invariant subspaces.

Proposition 2.1.6. *Given an invariant decomposition $V = V_1 \oplus V_2$ for the linear transformation $L : V \to V$, then the corresponding block structure for L is block diagonal. (I only used two subspaces here, but the case for more subspaces is identical.)*

Proof. Since $L(V_i) \subseteq V_i$, therefore for $i \neq j$, $p_j \circ L$ will kill V_i. So $L_{ij} = p_j \circ L \circ \iota_i = 0$. □

An eigen-direction is essentially a one-dimensional invariant subspace for our matrix. Since one dimensional subspace are spanned by a single vector, we sometimes just study eigenvectors. Finding a basis made of eigenvectors is essentially the same as finding a decomposition of V into invaraint one-dimensional subspaces. In particular, to block diagonalize a matrix is exactly the same as to find invariant decompositions of the domain.

Let us see a concrete example of this, using the same example as before.

Example 2.1.7. Consider a rotation in $\mathbb{R}^3$ around the line $x = y = z$ that sends the positive x-axis to the positive y-axis, and the positive y-axis to the positive z-axis, and the positive z-axis to the positive x-axis. We know its linear map has matrix $R = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. This matrix has non-real eigenvalues, so there is NO REAL diagonalizations. However, maybe we can find a REAL block-diagonalization?

There are two invariant subspaces that R must act on. One is the axis of rotation, the line $x = y = z$. This is a one-dimensional subspace V_1 spanned by $\begin{bmatrix}1\\1\\1\end{bmatrix}$. R acts on V_1 by simply fixing everyone, i.e., via the 1×1 matrix $R_{11} = \begin{bmatrix}1\end{bmatrix}$.

The other is the orthogonal complement of V_1, the subspace V_2 of all vectors $\begin{bmatrix}x\\y\\z\end{bmatrix}$ such that $x+y+z = 0$. Say we pick basis $\begin{bmatrix}1\\-1\\0\end{bmatrix}$ and $\begin{bmatrix}1\\0\\-1\end{bmatrix}$. Our linear map acts on V_2 as a rotation of $\frac{2\pi}{3}$, i.e., via some 2×2 matrix R_{22}. To find the matrix $R_{22} : V_2 \to V_2$, note that it depends on the basis we have chosen for V_2!!! So this is NOT going to be the standard rotation matrix, because we forgot to pick an orthonormal basis. Oops. Nevermind, let us just keep going forward.

Using the basis $\boldsymbol{v}_1 = \begin{bmatrix}1\\-1\\0\end{bmatrix}$ and $\boldsymbol{v}_2 = \begin{bmatrix}1\\0\\-1\end{bmatrix}$ for V_2, note that $R\boldsymbol{v}_1 = \begin{bmatrix}0\\1\\-1\end{bmatrix} = \boldsymbol{v}_2 - \boldsymbol{v}_1$, and $R\boldsymbol{v}_2 = \begin{bmatrix}-1\\1\\0\end{bmatrix} = -\boldsymbol{v}_1$. So $R_{22} = \begin{bmatrix}-1 & -1\\1 & 0\end{bmatrix}$.

So, under the basis $\begin{bmatrix}1\\1\\1\end{bmatrix}$, $\begin{bmatrix}1\\-1\\0\end{bmatrix}$ and $\begin{bmatrix}1\\0\\-1\end{bmatrix}$, our matrix will change into $\begin{bmatrix}R_{11} & \\ & R_{22}\end{bmatrix} = \begin{bmatrix}1 & & \\ & -1 & -1\\ & 1 & 0\end{bmatrix}$, which is block diagonal.

So we have $R = \begin{bmatrix}1 & 1 & 1\\1 & -1 & 0\\1 & 0 & -1\end{bmatrix}\begin{bmatrix}1 & & \\ & -1 & -1\\ & 1 & 0\end{bmatrix}\begin{bmatrix}1 & 1 & 1\\1 & -1 & 0\\1 & 0 & -1\end{bmatrix}^{-1}$.

Of course, as we can see in hind-sight, we can also find an orthonormal basis for V_2, say $\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\end{bmatrix}$ and $\frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\-2\end{bmatrix}$. Then R_{22} will be the standard rotation matrix $\begin{bmatrix}\cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3}\\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3}\end{bmatrix} = \begin{bmatrix}-\frac{1}{2} & -\frac{\sqrt{3}}{2}\\ \frac{\sqrt{3}}{2} & -\frac{1}{2}\end{bmatrix}$.

So we have$R = \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}}\end{bmatrix}\begin{bmatrix}1 & & \\ & -\frac{1}{2} & -\frac{\sqrt{3}}{2}\\ & \frac{\sqrt{3}}{2} & -\frac{1}{2}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}}\end{bmatrix}^{-1}$. We saved a bit of calculations but the numbers are uglier. Also note that the inverse here is also easy to calculate, because that matrix is now an orthogonal matrix, courtesy of picking an orthonormal basis. So the inverse here is just a transpose. In practise, this alone will make this better than the previous calculation, despite the ugly entries. ☺

Now, before we move on, let us consider the decompositions with more than two subspaces. These are mostly quoted from my linear algebra notes last semester.

Proposition 2.1.8. *For subspaces $V_1, \dots, V_k$ of a vector space V, the following are equivalent:*

1. *Pick any non-zero $\boldsymbol{v}_i \in V_i$ for each i, then $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is linearly independent.*
2. $\dim(\sum V_i) = \sum \dim V_i$.

Proof. We pick a basis $\mathcal{B}_i$ for each V_i. Let $\boldsymbol{v}_{i,j}$ be the j-th vector in $\mathcal{B}_i$. Let $\mathcal{B} = \bigcup_{i=1}^k \mathcal{B}_i$.

Forward Direction:

I claim that $\mathcal{B}$ is linearly independent, and then we are done.

To see this, suppose $\sum_{i,j} a_{i,j}\boldsymbol{v}_{i,j} = \mathbf{0}$. Then $\sum_i(\sum_j a_{i,j}\boldsymbol{v}_{i,j}) = \mathbf{0}$, but for each i, we can see that $\boldsymbol{w}_i = \sum_j a_{i,j}\boldsymbol{v}_{i,j} \in V_i$. Now $\sum_i \boldsymbol{w}_i = \mathbf{0}$, so the only possibility here is that all $\boldsymbol{w}_i = \mathbf{0}$.

Now for each i, $\sum_j a_{i,j}\boldsymbol{v}_{i,j} = \mathbf{0}$. But these $\boldsymbol{v}_{i,j}$ for fixed i form the basis $\mathcal{B}_i$, which is linearly independent. So all $a_{i,j}$ are zero.

Backward Direction:

If $\dim(\sum V_i) = \sum \dim V_i$, note that $\mathcal{B}$ must span $\dim(\sum V_i)$ and it has exactly $\sum \dim V_i$ vectors, and hence it must be a basis.

So if we pick any non-zero $\boldsymbol{v}_i \in V_i$ for each i, we have $\boldsymbol{v}_i = \sum_j a_{i,j}\boldsymbol{v}_{i,j}$ where some $a_{i,j} \neq 0$. If we have a linear combination $\sum_i b_i\boldsymbol{v}_i = \mathbf{0}$, then we have $\sum_{i,j} b_i a_{i,j}\boldsymbol{v}_{i,j} = \mathbf{0}$, which is a linear combination of vectors in $\mathcal{B}$. Hence all coefficients here are zero, and $b_i a_{i,j} = 0$ for all i, j.

For each i, since we must have some $a_{i,j} \neq 0$ for some j, it follows that $b_i = 0$. ☐

Let us redo the same proposition again, to see a slightly better proof.

Proposition 2.1.9. *(I like this proof a bit better because it avoids double indices.) For subspaces $V_1, \ldots, V_k$ of a vector space V, the following are equivalent:*

1. *Pick any non-zero $\boldsymbol{v}_i \in V_i$ for each i, then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is linearly independent.*
2. *For each i, then $(V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i = \{\mathbf{0}\}$.*
3. *$\dim(\sum V_i) = \sum \dim V_i$.*

Proof. ***(1) implies (2):***

Pick any $\boldsymbol{v} \in (V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i$. Then $\boldsymbol{v} = a_1\boldsymbol{v}_1 + \cdots + a_{i-1}\boldsymbol{v}_{i-1} + a_{i+1}\boldsymbol{v}_{i+1} + \cdots + a_k\boldsymbol{v}_k$, where we have non-zero $\boldsymbol{v}_j \in V_j$. Since we also have $\boldsymbol{v} \in V_i$, by setting $\boldsymbol{v}_i = \boldsymbol{v}$, we see that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ has a linear dependency! So we must have $\boldsymbol{v}_i = \mathbf{0}$.

(2) implies (3):

Note that by assumption, we have $(V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i = \{\mathbf{0}\}$ for each i. So we have $(V_1 + \cdots + V_{i-1}) \cap V_i = \{\mathbf{0}\}$ for each i as well.

By the inclusion-exclusion principle of subspace dimensions, we have $\dim(\sum V_i) = \dim(V_1 + \cdots + V_{k-1}) + \dim(V_k) - \dim((V_1 + \cdots + V_{k-1}) \cap V_k) = \dim(V_1 + \cdots + V_{k-1}) + \dim(V_k)$. Then we have $\dim(V_1 + \cdots + V_{k-1}) = \dim(V_1 + \cdots + V_{k-2}) + \dim V_{k-1} - \dim((V_1 + \cdots + V_{k-2}) \cap V_{k-1}) = \dim(V_1 + \cdots + V_{k-2}) + \dim V_{k-1}$, and so on. Thus inductively we have $\dim(\sum V_i) = \sum \dim V_i$.

(3) implies (2):

We do induction. If $k = 1$, there is nothing to prove. Suppose $k > 1$.

For each i, note that $\dim(\sum V_i) = \dim(V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) + \dim V_i - \dim((V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i) \leq \dim(V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) + \dim V_i \leq \sum \dim V_i = \dim(\sum V_i)$. So we have equality everywhere, and in particular we must have $\dim((V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i) = 0$.

(2) implies (1):

Pick any non-zero $\boldsymbol{v}_i \in V_i$ for each i. Suppose we have a linear dependency $\sum a_i\boldsymbol{v}_i = \mathbf{0}$. If $a_i \neq 0$, then $\boldsymbol{v}_i$ will be a linear combination of the other vectors, i.e., $\boldsymbol{v}_i \in (V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k) \cap V_i$. So we must have $a_i = 0$. ☐

If either of these two conditions is satisfied, then we say the subspaces $V_1, \ldots, V_k$ are linearly independent. Keep in mind that pairwise independence does NOT imply collective independence. Consider the following example.

Example 2.1.10. Let U, V, W be three subspaces of $\mathbb{R}^2$ such that U is the x-axis, V is the y-axis, and W is the line defined by the equation $x = y$. Then note that U, V, W are pairwise independent, but collectively, they are NOT linearly independent.

This counter example is important to keep in mind. For example, subset algebra satisfies the law of distribution. (I.e., in set theory, $S_1 \cap (S_2 \cup S_3) = (S_1 \cap S_2) \cup (S_1 \cap S_3)$ and $S_1 \cup (S_2 \cap S_3) = (S_1 \cup S_2) \cap (S_1 \cup S_3)$ for any three subsets.) However, subspace algebra does NOT have the law of distribution. You can verify that, in our example, $U \cap (V + W) \neq (U + V) \cap (U + W)$ and similarly $U + (V \cap W) \neq (U + V) \cap (U + W)$.

This is also closely related to probability theory. For many random variables, pairwise independently distributed does NOT imply collectively independently distributed. And the counter example there is essentially a modified version of our example here. (Just change our field $\mathbb{R}$ into any finite field, and build variables X, Y, Z whose distribution is defined via the subspaces U, V, W.) ☺

In a similar manner as before, block diagonalizations are related to invariant decomposition of the domain $\mathbb{R}^n$ into a direct sum of linearly independent subspaces.

We end this with a quick lemma for future use.

2.1.3 Searching for good invariant decomposition

So this is it. How can we find a good invariant decomposition? Let us first see what kinds of invariant subspaces we have.

Example 2.1.11. Given any matrix A, consider the zero space $\mathrm{Ker}(A)$. obviosuly $A(\mathrm{Ker}(A)) = \{0\} \subseteq \mathrm{Ker}(A)$. So this is indeed an invariant subspace!

Dually, since A sends everything into $\mathrm{Ran}(A)$ by definition, we have $A(\mathrm{Ran}(A)) \subseteq \mathrm{Ran}(A)$ as well. Hooray! Another invariant subspace!

In fact, for $n \times n$ matrices A, we also have $\dim \mathrm{Ker}(A) + \dim \mathrm{Ran}(A) = n$. This is a really good omen. In fact, consider say $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. Then $\mathrm{Ker}(A)$ and $\mathrm{Ran}(A)$ are both invaraint subspaces, and in fact we have $\mathbb{R}^3 = \mathrm{Ker}(A) \oplus \mathrm{Ran}(A)$ in this case, a perfect decomposition into invariant subspaces!

Unfortunately, we do not always have this. Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then $\mathrm{Ker}(A) = \mathrm{Ran}(A)$. So we failed in this case.

In fact, the best complement subspace for $\mathrm{Ker}(A)$ is actually $\mathrm{Ran}(A^{\mathrm{T}})$ (or $\mathrm{Ran}(A^*)$ in the complex case), and we always have $\mathbb{R}^n = \mathrm{Ker}(A) \oplus \mathrm{Ran}(A^{\mathrm{T}})$. However, again consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, you shall see that $\mathrm{Ran}(A^{\mathrm{T}})$ is usually not an invaraint subspace!

We are screwed either way. ☺

What can we do then? Well, recall our original motivation of doing diagonalization. What started us on this path about eigenstuff and diagonalization? The original motivation is to understand iterated applications of the same matrix, i.e., the eventual behavior of the sequence $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^n\boldsymbol{v}, \ldots$. Diagonalization gives us a quick way to calculate A^n for large n.

As a result, maybe we shouldn't focus on the *immediate* kernel and range of A. Rather, we should focus on the *eventual* kernel and range of A.

Example 2.1.12. Consider $A = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$. Then applying A repeatedly, we have:

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \xmapsto{A} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \xmapsto{A} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \xmapsto{A} \mathbf{0}.$$

Then we say $\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ is *eventually* killed by A. Let N_∞ be the subspace of all vectors eventually killed by A.

Also note that $A^2 = \begin{bmatrix} 0 & 0 & 1 & \\ & 0 & 0 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$ and $A^n = \begin{bmatrix} 0 & 0 & 0 & \\ & 0 & 0 & \\ & & 0 & \\ & & & 1 \end{bmatrix}$ for all $n \geq 3$. So eventually, $A^n\boldsymbol{v}$ will be a multiple of $\boldsymbol{e}_4$ for large enough n. So we say the *eventual* range of A is the subspace R_∞ spanned by $\boldsymbol{e}_4$.

Check yourself that in fact $\mathbb{R}^4 = N_\infty \oplus R_\infty$ is an invariant decomposition. ☺

Definition 2.1.13. *Given a linear map or a matrix A, we define $N_\infty(A) = \cup_{k=1}^\infty \operatorname{Ker}(A^k)$ and $R_\infty(A) = \cap_{k=1}^\infty \operatorname{Ran}(A^k)$.*

In particular, $\boldsymbol{v} \in N_\infty(A)$ if and only if some powers of A will kill $\boldsymbol{v}$. And $\boldsymbol{v} \in R_\infty(A)$ if and only if $\boldsymbol{v}$ is in the range of ALL powers of A.

It turns out that we don't really have to look at all powers of A. Whatever A kills, then A^2 must kill as well. So as k grows, the subspace $\operatorname{Ker}(A^k)$ will be non-decreasing. However, its dimension is at most n (the dimension of the domain). So it cannot grow forever, and eventually it must stabilize. So we see that $N_\infty(A) = \operatorname{Ker}(A^k)$ for some k. We in fact have more. It turns out that k does not need to be too large.

Proposition 2.1.14. *For any $n \times n$ matrix A, we have $N_\infty(A) = \operatorname{Ker}(A^k)$ for some $k \leq n$. (In particular, we always have $N_\infty(A) = \operatorname{Ker}(A^n)$.)*

Proof. Let k be the smallest integer such that $A^k\boldsymbol{v} = \boldsymbol{0}$. Then by the lemma below, $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{k-1}\boldsymbol{v}$ are linearly independent. But now we have k linearly independent vectors in $\mathbb{R}^n$, so $k \leq n$. □

Let us prove this lemma here. It claims that for a killing chain $\boldsymbol{v} \overset{A}{\mapsto} A\boldsymbol{v} \overset{A}{\mapsto} \ldots \overset{A}{\mapsto} A^{k-1}\boldsymbol{v} \overset{A}{\mapsto} \boldsymbol{0}$, everything will be independent before $\boldsymbol{v}$ is finally killed.

Lemma 2.1.15. *For any $n \times n$ matrix A, and any $\boldsymbol{v} \in N_\infty(A)$, let k be the smallest integer such that $A^k\boldsymbol{v} = \boldsymbol{0}$. Then $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{k-1}\boldsymbol{v}$ are linearly independent.*

Proof. (As an illustrative example, say wehave $k = 4$, so $A^4\boldsymbol{v} = \boldsymbol{0}$. Suppose for contradiction, say we have a linear relation $3A\boldsymbol{v} + 2A^2\boldsymbol{v} + 4A^3\boldsymbol{v} = \boldsymbol{0}$. Then multiply A^2 to both sides, we have $\boldsymbol{0} = 3A^3\boldsymbol{v} + 2A^4\boldsymbol{v} + 4A^5\boldsymbol{v} = 3A^3\boldsymbol{v}$. So $A^3\boldsymbol{v} = \boldsymbol{0}$. Contradiction indeed.)

Suppose we have a nontrivial relation $\sum_{i=0}^{k-1} a_i A^i\boldsymbol{v} = \boldsymbol{0}$.Let j be the smallest natural number such that $a_j \neq 0$. Then multiply A^{k-j-1} on both sides of $\sum_{i=0}^{k-1} a_i A^i\boldsymbol{v} = \boldsymbol{0}$, and use the fact that $A^k\boldsymbol{v} = \boldsymbol{0}$, we see that $a_j A^{k-1}\boldsymbol{v} = \boldsymbol{0}$. Then since $a_j \neq 0$, we see that $A^{k-1}\boldsymbol{v} = \boldsymbol{0}$. Contradiction.

So all linear relations among $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{k-1}\boldsymbol{v}$ are trivial. These vectors are linearly independent. □

As you can see, vectors should be your role models. I hope that after college, you shall grow into an independent person until you die, like these vectors here.

We also have a similar result for the "eventual range" of A.

Proposition 2.1.16. $N_\infty(A) = \operatorname{Ker}(A^k)$ *if and only if* $R_\infty(A) = \operatorname{Ran}(A^k)$.

Proof. Note that as k increases, $\operatorname{Ran}(A^k)$ is a non-increasing chain of subspaces. But since $\dim \operatorname{Ran}(A^k) = n - \dim \operatorname{Ker}(A^k)$, we see that $\dim \operatorname{Ran}(A^k)$ must stabilize as soon as $\dim \operatorname{Ker}(A^k)$ stabilizes, and hence that $\operatorname{Ran}(A^k)$ must stabilize as soon as $\operatorname{Ker}(A^k)$ stabilizes. □

Let us now show that we indeed have invariant subspaces.

Proposition 2.1.17. *For any polynomial $p(x)$, then $\operatorname{Ker}(p(A))$ and $\operatorname{Ran}(p(A))$ are A-invariant.*

Proof. The key is the fact that $xp(x) = p(x)x$ as polynomials. As a result, $Ap(A) = p(A)A$ as matrices because they are the same polynomial of A.

Suppose $p(A)\boldsymbol{v} = \boldsymbol{0}$. Then $p(A)(A\boldsymbol{v}) = p(A)A\boldsymbol{v} = Ap(A)\boldsymbol{v} = A(\boldsymbol{0}) = \boldsymbol{0}$. So $\operatorname{Ker}(p(A))$ is A-invariant.

Suppose $\boldsymbol{v} = p(A)\boldsymbol{w}$ for some $\boldsymbol{w}$. Then $A\boldsymbol{v} = Ap(A)\boldsymbol{w} = p(A)(A\boldsymbol{w})$. So $\operatorname{Ran}(p(A))$ is A-invariant. □

Corollary 2.1.18. $N_\infty(A)$ *and* $R_\infty(A)$ *are A-invariant.*

Theorem 2.1.19 (The Ultimate Invariant Decomposition)**.** *For any $n \times n$ matrix A, we have an invariant decomposition $\mathbb{R}^n = N_\infty(A) \oplus R_\infty(A)$.*

Proof. We already know that these two are invariant subspaces. Also, since for some $k \leq n$ we have $N_\infty(A) = \mathrm{Ker}(A^k)$ and $R_\infty(A) = \mathrm{Ran}(A^k)$, therefore we have $\dim N_\infty(A) + \dim R_\infty(A) = n$. So we only need to show that they have zero intersection.

(Remark: For a collection of vectors, having n vectors, linearly independent, spanning, any two of these three conditions would imlpy that we have a basis. In a comparative manner, dimensions add up to n, zero intersection, sum space is the whole space, any two of these three conditions would imlpy that we have a direct sum.)

Suppose $\boldsymbol{v} \in N_\infty(A) \cap R_\infty(A)$. Since $\boldsymbol{v} \in N_\infty(A)$, we have some $k \leq n$ such that $A^k\boldsymbol{v} = \mathbf{0}$. But since $\boldsymbol{v} \in R_\infty(A) \subseteq \mathrm{Ran}(A^n)$, we have $\boldsymbol{v} = A^n\boldsymbol{w}$ for some $\boldsymbol{w}$. Then $A^{k+n}\boldsymbol{w} = \mathbf{0}$, so $\boldsymbol{w} \in N_\infty(A)$ as well. But this implies that $\boldsymbol{w} \in \mathrm{Ker}(A^n)$, and hence $\boldsymbol{v} = A^n\boldsymbol{w} = \mathbf{0}$. Oops. So we are done.

(Essentially, the key idea is that $N_\infty(A)$ stabilizes after finitely many steps, while $\boldsymbol{v} \in R_\infty(A)$ means we can realize $\boldsymbol{v}$ after arbitrarily many steps, which forces $\boldsymbol{v} \in N_\infty(A)$ to be zero.) ☐

2.1.4 (Review) Polynomials of Matrices

It has come to my attention that some of our classmates have never seen this. So let us do it here as a review. Note that everything in this section could be over $\mathbb{R}$ or over $\mathbb{C}$, it does not matter much.

Remark 2.1.20. *This remark is not necessary. Feel free to skip this remark entirely.*

Let us define what a polynomial is.

We define a real (or complex) polynomial $p(x)$ toi be a finite sequence of real (or complex) numbers, say $(a_0, \ldots, a_n)$. We also write $p(x) = a_0 + a_1x + \cdots + a_nx^n$ where the symbol x^k has no specific meaning, and it is simply a place holder.

We add polynomial such that $(a_0, \ldots, a_n) + (b_0, \ldots, b_m) = (a_0 + b_0, \ldots, a_n + b_n, b_{n+1}, \ldots, b_m)$ if $m > n$. We multiply polynomial such that $(a_0, \ldots, a_n)(b_0, \ldots, b_m) = (c_0, \ldots, c_{m+n})$ where $c_k = \sum_{i=0}^k a_ib_{k-i}$.

Now, see if you can prove the following:

All polynomials form a vector space V, with a basis $1, x, x^2, \ldots$. For any bilinear map $m : V \times V \to V$ such that $m(x^a, x^b) = x^{a+b}$, then m must be the polynomial multiplication as in our definition.

You do NOT need to remember the formula, or worry about this definition. I want you to see this definition NOT because it is useful. It is not. Writing $p(x) = 4 + 2x + 3x^2$ is strictly better than writing $(4, 2, 3)$.

However, this definition makes clear of the fact that a polynomial does NOT need x to have any meaning. It could be a real number, a complex number, a matrix, a whatever. We can give whatever meaning to x, and as long as x is capable of having a "power structure", then we can define $p(x)$ accordingly as the linear combination of corresponding powers.

Here by power structure, it means that we want x^k to be defined, and we want the property that $x^ax^b = x^{a+b}$.

What is a polynomial, say $p(x) = 4 + 2x + 3x^2$? Well, in the realm of linear algebra, the best answer is that "a polynomial is a linear combination of powers." In our case, $p(x)$ is a linear combination of $1, x, x^2$. (Note that $1 = x^0$, if you like.)

For each square matrix A, we obviously have well-defined powers of A. Therefore, if $p(x)$ is some linear combination of powers of x, we can define $p(A)$ to be the corresponding linear combination of powers of A. Easy peasy.

Proposition 2.1.21. *For any polynomials $p(x), q(x)$, and any square matrix A, then $p(A)+q(A) = (p+q)(A)$ and $p(A)q(A) = (pq)(A)$. (Here $(p + q)(x)$ is the polynomial $p(x) + q(x)$ and $(pq)(x)$ is the polynomial $p(x)q(x)$.)*

Proof. DIY. ☐

Now, why do we study polynomials of matrices? It is mainly because powers A^k has many good properties related to A, and thus linear combinations of these powers, $p(A)$, would also share such properties. Here let us write some.

Proposition 2.1.22. *For any polynomials $p(x), q(x)$, we have $p(A)q(A) = q(A)p(A)$.*

Proof. First, note that $AA^k = A^{k+1} = A^kA$. Therefore A commutes with powers of A. Therefore A commutes with linear combinations of powers of A, i.e., polynomials of A.

So $p(A)$ commutes with A. Therefore $p(A)$ commutes with powers of A. Therefore $p(A)$ commutes with linear combinations of powers of A, i.e., other polynomials of A, say $q(A)$. So $p(A)q(A) = q(A)p(A)$. $\square$

We also have good results about eigenstuff.

Proposition 2.1.23. *$A\boldsymbol{v} = \lambda\boldsymbol{v}$ implies that $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v}$.*

Proof. If $A\boldsymbol{v} = \lambda\boldsymbol{v}$, then it is easy to see that $A^k\boldsymbol{v} = \lambda^k\boldsymbol{v}$. Now we take linear combinations of various powers, we see that $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v}$. $\square$

Corollary 2.1.24. *If A has eigenvalues $\lambda_1, \ldots, \lambda_n$ counting algebraic multiplicity, then $p(A)$ has eigenvalues $p(\lambda_1), \ldots, p(\lambda_n)$ counting algebraic multiplicity. And each eigenvector of A for some eigenvalue λ is an eigenvector of $p(A)$ for the eigenvalue $p(\lambda)$.*

Now, the eigenvectors of A are all eigenvectors of $p(A)$, but sometimes $p(A)$ has other eigenvectors.

Example 2.1.25. Consider $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Its eigenvectors are vectors on the coordinate-axes. But $A^2 = I$, so ALL vectors are eigenvectors of A^2. As you can see, this is because distinct eigenvalues of A are collapsed into the same eigenvalue of $p(A)$.

Also consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Its eigenvectors are vectors on the x-axis. But $A^2 = O$, so ALL vectors are eigenvectors of A^2. As you can see, this is because A cannot be diagonalized (non-trivial Jordan block....), yet A^2 kills the obstruction to diagonalization (chopped down the bad Jordan block into smaller blocks, i.e., 1×1 blocks), so now A^2 CAN be diagonalized. ☺

So how can we find all eigenvectors of $p(A)$? Under some special cases, the answers are easy.

Proposition 2.1.26. *Suppose A can be diagonalized. Pick any polynomial $p(x)$. For any eigenvalue λ of $p(A)$, let $\lambda_1, \ldots, \lambda_k$ be all eigenvalues of A such that $p(\lambda_i) = \lambda$. Then $p(A)\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $\boldsymbol{v}$ is a linear combination of eigenvectors of A for the eigenvalues $\lambda_1, \ldots, \lambda_k$.*

Proof. Diagonalize $A = BDB^{-1}$. Then $p(A) = Bp(D)B^{-1}$. So up to a change of basis, we can assume that A is diagonal. Then DIY. $\square$

We can have more results if we delve into theory of polynomials. The following are entirely optional. Read on if you like.

Example 2.1.27. Skip this example if you know about the Euclidean algorithm for coprime integers. Otherwise, read on.

Consider 22 and 15. They have no common prime factor. They are coprime.

We divide 22 by 15, and we shall get a remainder. We have $22 = 15 + 7$. Next we divide 15 by 7 and get $15 = 7 \times 2 + 1$. So we eventually reduced to the remainder 1.

Putting these together, we have $1 = 15 - 2 \times 7 = 15 - 2 \times (22 - 15) = 3 \times 15 - 2 \times 22$. So an integer-linear combination of 15 and 22 gives 1. This process is called the Euclidean algorithm, and it shows that two numbers x, y are coprime if and only if we can find integers a, b such that $ax + by = 1$.

Now we do the same thing for polynomials. Note that the polynomial $p(x) = x^3 + 3x^2 + 3x + 1$ and $q(x) = x^2 - 3x + 2$ has no common root, i.e., upon factorization, they shall have no common non-constant factor. They are coprime polynomials.

We divide x^3+3x^2+3x+1 by x^2-3x+2, and we shall get a remainder. We have $x^3+3x^2+3x+1 = (x^2-3x+2)(x+6)+(19x-11)$. Next we divide x^2-3x+2 by $19x-11$, and we have $x^2-3x+2 = (19x-11)(\frac{1}{19}x+\frac{46}{19})+\frac{544}{19}$. So we eventually reduced to a constant remainder $\frac{544}{19}$.

Putting these together, we have $1 = \frac{19}{544}\frac{544}{19} = \frac{19}{544}((x^2-3x+2)-(19x-11)(\frac{1}{19}x+\frac{46}{19})) = \frac{19}{544}((x^2-3x+2)-(\frac{1}{19}x+\frac{46}{19})((x^3+3x^2+3x+1)-(x^2-3x+2)(x+6)))$. Break down the parenthesis, we see that we can find polynomials $a(x), b(x)$ such that $a(x)p(x)+b(x)q(x)=1$. ☺

Theorem 2.1.28. *If two complex polynomial $p(x), q(x)$ has no common root, then we can find polynomials $a(x), b(x)$ such that $a(x)p(x)+b(x)q(x)=1$.*

Proof. Outside the scope of this class. Search for Euclidean algorithm online. □

Corollary 2.1.29. *If two complex polynomial $p(x), q(x)$ has no common root, then for any square matrix A, $\mathrm{Ker}(p(A)q(A)) = \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$.*

Proof. Since $p(x), q(x)$ has no common root, we can find polynomials $a(x), b(x)$ such that $a(x)p(x)+b(x)q(x)=1$. Then $a(A)p(A)+b(A)q(A)=I$.

Suppose $\boldsymbol{v} \in \mathrm{Ker}(p(A)) \cap \mathrm{Ker}(q(A))$. Then $p(A)\boldsymbol{v} = \mathbf{0}$ and $q(A)\boldsymbol{v} = \mathbf{0}$. Then $\boldsymbol{v} = I\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} = \mathbf{0}$. So we have trivial intersection.

Next, if $\boldsymbol{v} \in \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$, then $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ where $p(A)\boldsymbol{v}_1 = \mathbf{0}$ and $q(A)\boldsymbol{v}_2 = \mathbf{0}$. Then $p(A)q(A)\boldsymbol{v} = q(A)p(A)\boldsymbol{v}_1 + p(A)q(A)\boldsymbol{v}_2 = \mathbf{0}+\mathbf{0} = \mathbf{0}$. So we see that $\mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A)) \subseteq \mathrm{Ker}(p(A)q(A))$.

Conversely, suppose $\boldsymbol{v} \in \mathrm{Ker}(p(A)q(A))$. Then $a(A)p(A)\boldsymbol{v} \subseteq \mathrm{Ker}(q(A))$ and $b(A)q(A)\boldsymbol{v} \subseteq \mathrm{Ker}(p(A))$. Then we have $\boldsymbol{v} = I\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} \in \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$. So we have $\mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A)) \supseteq \mathrm{Ker}(p(A)q(A))$. □

Corollary 2.1.30. *Suppose $p(x)$ has distinct roots. Pick any square matrix A. For any eigenvalue λ of $p(A)$, let $\lambda_1, \dots, \lambda_k$ be all eigenvalues of A such that $p(\lambda_i) = \lambda$. Then $p(A)\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $\boldsymbol{v}$ is a linear combination of eigenvectors of A for the eigenvalues $\lambda_1, \dots, \lambda_k$.*

Proof. Replace A by $A-\lambda I$ if needed, we can WLOG say $\lambda = 0$. Then $p(A)\boldsymbol{v} = \mathbf{0}$ implies that $\boldsymbol{v}$ is a linear combination of vectors $\boldsymbol{v}_i \in \mathrm{Ker}(A-\lambda_i I)$. So we are done.

The converse direction is trivial. □

Corollary 2.1.31. *If $p(x)$ has distinct roots $\lambda_1, \dots, \lambda_n$, then the solutions to the differential equation $p(\frac{d}{dx})f = 0$ are linear combinations of $\mathrm{e}^{\lambda_i x}$.*

Proof. Taking derivative $\frac{d}{dx}$ is a linear operation, and for any complex number λ, $\frac{d}{dx}$ has eigenvalue λ with eigenvectors multiples of $\mathrm{e}^{\lambda x}$. So we are done. □

Example 2.1.32. Consider an object attached to a spring, and it is bouncing around horizontally without friction. Say the elastic coefficient is 1, object mass is 1, and the location of our object at time t is $f(t)$. Then $f''(t) = -f(t)$.

So let $p(x) = x^2+1$, we have $p(\frac{d}{dx})f = 0$. Note that $p(x)$ has distinct roots, so the solutions are linear combinations of $\mathrm{e}^{\mathrm{i}t}$ and $\mathrm{e}^{-\mathrm{i}t}$. Taking real solutions only, then we see that the solutions are linear combinations of $\sin t$ and $\cos t$.

So our object moves periodically.

If we have elastic coefficient k, and say we have friction positively correlated to speed with coefficient μ, and object mass m. Then $mf''(t) = -kf(t) - \mu f'(t)$. So let $p(x) = mx^2+\mu x+k$, and we have $p(\frac{d}{dx})f = 0$ again. Hopefully we have distinct roots (which we almost always have), then we are good to go again. ☺

2.1.5 Generalized Eigenspace

In our previous sections, we have been doing linear algebra over $\mathbb{R}$. But it is just the same over $\mathbb{C}$. For the rest of the section, we are restricting our attention to $\mathbb{C}$ because we need those eigenvalues.

Remark 2.1.33. *Usually, things done in $\mathbb{R}$ are easily true over $\mathbb{C}$ (as long as no inner product is involved), but things done in $\mathbb{C}$ might NOT be true over $\mathbb{R}$. For example, any $n \times n$ matrix over $\mathbb{C}$ has n eigenvalues in $\mathbb{C}$ counting algebraic multiplicity. But the statement is NOT true if we replace $\mathbb{C}$ by $\mathbb{R}$.*

Our goal here is the following. For any matrix A, we aim to block diagonalize it, such that each diagonal block is a matrix with all eigenvalues the same. For example, something like this: $\begin{bmatrix} 1 & 2 & 3 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$. Here there are two diagonal blocks, the first one has all eigenvalues 1, and the second one has all eigenvalues 2.

In essense, we are looking for an invariant decomposition $\mathbb{C}^n = V_1 \oplus \cdots \oplus V_k$ such that A retricted to each V_i will be a matrix with all eigenvalues the same.

Our previous ultimate invariant decomposition is already in this direction. Suppose $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ as the corresponding block-diagonalization of A for the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. Now, A_N is the restriction of A to a linear transformation on $N_\infty(A)$, and it will eventually kill everything in this domain, so A_N can only have zero eigenvalues.

In contrast, Since $\mathrm{Ker}(A) \subseteq N_\infty(A)$ and $N_\infty(A) \cap R_\infty(A) = \{\mathbf{0}\}$, it turns out that A restricted to a linear transformation on $R_\infty(A)$ will have zero kernel, i.e., A_R is an invertible matrix! So it has no zero eigenvalue.

In particular, the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$ has successfully isolated all the zero-eigenvalue behaviors of A in $N_\infty(A)$, and all the non-zero-eigenvalue behaviors of A to $R_\infty(A)$.

Recall that the eigenspace of a matrix A for the eigenvalue λ is simply $\mathrm{Ker}(A - \lambda I)$. We now define the following.

Definition 2.1.34. *The generalized eigenspace of a matrix A for the eigenvalue λ is the subspace $N_\infty(A-\lambda I)$.*

Let us show that these subspaces are linearly independent.

Lemma 2.1.35. *If $\lambda \neq \mu$, then $N_\infty(A - \lambda I) \subseteq R_\infty(A - \mu I)$.*

Proof. Replace A by $A - \mu I$ if needed, it is enough to prove that $N_\infty(A - \lambda I) \subseteq R_\infty(A)$ whenever $\lambda \neq 0$.

Pick any $\boldsymbol{v} \in N_\infty(A - \lambda I) = \mathrm{Ker}(A - \lambda I)^n$. Our goal is to show that $\boldsymbol{v} \in \mathrm{Ran}(A^k)$ for all k. We have $(A - \lambda I)^n \boldsymbol{v} = \mathbf{0}$. Expanding this, since $\lambda \neq 0$, on the left hand side we have something like $A(\text{stuff})\boldsymbol{v} + (\text{non-zero constant})\boldsymbol{v} = \mathbf{0}$, which can be rearranged into $\boldsymbol{v} = A(\text{stuff})\boldsymbol{v}$, and its iteration shall give us the result. And we are done.

More formally, let $(x - \lambda)^n = xp(x) + (-\lambda)^n$ for some polynomial $p(x)$. So $\mathbf{0} = (A - \lambda I)^n \boldsymbol{v} = Ap(A)\boldsymbol{v} + (-\lambda)^n \boldsymbol{v}$. Let $B = -\frac{1}{(-\lambda)^n} p(A)$, we see that $\boldsymbol{v} = AB\boldsymbol{v}$ where $AB = BA$. Then it is easy to see that $\boldsymbol{v} = ABAB\boldsymbol{v} = A^2 B^2 \boldsymbol{v}$ and so on. So $\boldsymbol{v} = A^k B^k \boldsymbol{v} \in \mathrm{Ran}(A^k)$ for all k. So $\boldsymbol{v} \in \cap \mathrm{Ran}(A^k) = R_\infty(A)$. ☐

Note that this immediately implies independence.

Corollary 2.1.36. *Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of A (NOT counting algebraic multiplicity, i.e., they are distinct complex numbers). Let $V_i = N_\infty(A - \lambda_i I)$ be the generalized eigenspace for each i. Then $V_1, \ldots, V_k$ are linearly independent subspaces ,and they are invariant under A.*

Proof. We need to show that $N_\infty(A - \lambda_i I)$ and $\bigcup_{j \neq i} N_\infty(A - \lambda_j I)$ have zero intersection. Note that we have $N_\infty(A - \lambda_i I) \cap R_\infty(A - \lambda_i I) = \{\mathbf{0}\}$, so it is enough to know that $N_\infty(A - \lambda_j I) \subseteq R_\infty(A - \lambda_i I)$ whenever $j \neq i$. And this is just the last lemma. ☐

They are not only independent. They in fact gives us the desired invariant decomposition of the whole domain.

Proposition 2.1.37 (Geometric meaning of algebraic multiplicity)**.** *Let λ be an eigenvalue of a square matrix A with algebraic multiplicity m, and let $V_\lambda = N_\infty(A - \lambda I)$ be the generalized eigenspace. Then* $\dim V_\lambda = m$.

Proof. Replacing A by $A - \lambda I$ if necessary, we can assume that $\lambda = 0$.

Now let $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ be the corresponding block diagonalization of A after a change of basis according to the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. As we have discussed before, A_N will only have eigenvalue zero, while A_R has no zero eigenvalue. But their characteristic polynomials must satisfy $p_A(x) = p_{A_N}(x)p_{A_R}(x)$. So the algebraic multiplicity of 0 in p_A is exactly the same as the degree of p_{A_N}, whcih is $\dim N_\infty(A)$. $\square$

Theorem 2.1.38. *Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of A (NOT counting algebraic multiplicity, i.e., they are distinct complex numbers). Let $V_i = N_\infty(A - \lambda_i I)$ be the generalized eigenspace for each i. Then we have an invariant decomposition $\mathbb{C}^n = \bigoplus_{i=1}^k V_i$.*

Proof. These subspaces are linearly independent, and their dimensions add up to n (since algebraic multiplicities add up to n). $\square$

Recall that previously, we see that all eigenvalues of A_N must be zero in the block diagonalization $\begin{bmatrix} A_N & O \\ O & A_R \end{bmatrix}$ corresponding to the invariant decomposition $\mathbb{C}^n = N_\infty(A) \oplus R_\infty(A)$. Similarly, given a block diagonalization of A, say $\begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}$ according to the generalized eigenspaces, then each A_i is the restriction of A to V_i, so all eigenvalues of A_i must be λ_i.

2.2 Nilpotent Matrices

2.2.1 Invariant Filtration and Triangularization

We have now block diagonalized our matrix, where each block is a matrix whose eigenvalues are all the same. What now? Well, we need to understand such matrices whose eigenvalues are all the same! Let us start with a special case. What if all eigenvalues are zero?

Definition 2.2.1. *We say a matrix A is nilpotent if $A^k = O$ for some positive integer k. (I.e., $N_\infty(A)$ is the whole domain.)*

(Tiny remark: "nil" means zero. "potent" means power. "Some power is zero", i.e., nilpotent.)

Remark 2.2.2. *If $A^k = O$ for some positive integer k, then we can in fact require that $k \leq n$. This is because of our previous analysis of $N_\infty(A)$. In particular, we always have $A^n = O$.*

Proposition 2.2.3. *A is nilpotent if and only if all eigenvalues of A are zero.*

Proof. Suppose A is nilpotent.

If A has eigenvalues $\lambda_1, \ldots, \lambda_n$ counting algebraic multiplicity, then $p(A)$ has eigenvalues $p(\lambda_1), \ldots, p(\lambda_n)$ counting algebraic multiplicity for any polynomial $p(x)$.

Now A^k has eigenvalues $\lambda_1^k, \ldots, \lambda_n^k$. Yet all eigenvalues of A^k are zero. Done.

Now suppose all eigenvalues of A^k are zero. Then since the domain is the direct sum of generalized eigenspaces, and A^k only has zero eigenvalue, hence $N_\infty(A)$ is the entire domain. So we are done. $\square$

Now, these nilpotent matrices are annoying. Many of them has NO good invariant decomposition at all! Instead, they behave like onions: layers of invariant subspaces, each containing the next.

Example 2.2.4. Consider $A = \begin{bmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{bmatrix}$. This is the "shift up" operator that sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} y \\ z \\ 0 \end{bmatrix}$, i.e., it is shifting the coordinates upwards. Therefore, we obviously have $A^3 = O$. It is nilpotent.

Now what are its invariant subspaces? If A is invariant, and $A(V) \subseteq V$ for some subspace V, then A restricted to this linear tranformation on V would be nilpotent as well. Now if $\dim V = k$, any nilpotent linear transformation must die in k steps. So we must have $A^k(V) = \{\mathbf{0}\}$.

(Alternatively, since $A^n = O$, consider the sequence of subspaces $V, A(V), \dots, A^n(V)$, then this sequence must eventually shrink to zero. Now if $A^i(V) = A^{i+1}(V)$, then $A^{i+2}(V) = A(A^{i+1}(V)) = A(A^i(V)) = A^{i+1}(V) = A^i(V)$, and the sequence would stabilize forever. So this sequence must shrink strictly until it hit zero. Each step the dimension must reduce by at least one. So if $\dim V = k$, we must have $A^k(V) = \{\mathbf{0}\}$.)

So $V \subseteq \operatorname{Ker}(A^k)$. However, in our case, note that for any k, $\operatorname{Ker}(A^k)$ is spanned by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_k$. So $\dim \operatorname{Ker}(A^k) = k = \dim V$, wow! So $V = \operatorname{Ker}(A^k)$.

In particular, all invariant subspaces of A are $\operatorname{Ker}(A^k)$ for some k. The invariant subspaces are exactly $\{0\}$, x-axis, xy-plane, and the whole space.

There is no invariant decomposition of the whole domain other than the trivial one. However, you can see that these invariant subspaces come in layers, like an onion, each layer containing the last. Why are Jordan blocks like $\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$? As we shall see later, it is precisely due to this onion structure. ☺

Definition 2.2.5. *Given a vector space V, a filtration for V is a sequence of subspaces $V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n = V$, where $\dim V_k = k$. For any linear tranformation $A : V \to V$, we say this is an (A-)invariant filtration if all V_k are A-invariant subspaces.*

So the idea is this: invariant decomposition leads to block diagonalization. Invariant filtration would lead to triangularization.

Proposition 2.2.6. *If $L : V \to V$ is a linear transformation, and V has an invariant filtration $V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n = V$. Pick any $\boldsymbol{v}_i \in V_i - V_{i-1}$ for each $1 \leq i \leq n$, then $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ form a basis of V, under which the matrix for L is upper triangular.*

Proof. Let us first show that $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ form a basis. It is enough to show linear independence.

We perform induction. Since $\boldsymbol{v}_1 \in V_1 - V_0$, it is non-zero, so it is linearly independent. For each $i \geq 1$, $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1} \in V_{i-1}$, yet $\boldsymbol{v}_i \notin V_{i-1}$. By induction hypothesis, $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}$ are already linearly independent, so $\boldsymbol{v}_1, \dots, \boldsymbol{v}_i$ are linearly independent as well. We are done.

In fact, it is not hard to see that $\boldsymbol{v}_1, \dots, \boldsymbol{v}_i$ form a basis for V_i for each i.

Now $\boldsymbol{v}_i \in V_i$, so by invariance, $L\boldsymbol{v}_i \in V_i$ as well. Say $L\boldsymbol{v}_i = a_{1i}\boldsymbol{v}_1 + \cdots + a_{ii}\boldsymbol{v}_i$ since $\boldsymbol{v}_1, \dots, \boldsymbol{v}_i$ form a basis for V_i.

Now by straight forward calculation, we have:

$$L(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = (a_{11}\boldsymbol{v}_1, a_{12}\boldsymbol{v}_1 + a_{22}\boldsymbol{v}_2, \dots, a_{1n}\boldsymbol{v}_1 + \cdots + a_{nn}\boldsymbol{v}_n) = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \begin{bmatrix} a_{11} & \dots & a_{1n} \\ & \ddots & \vdots \\ & & a_{nn} \end{bmatrix}.$$

This means that using $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ as basis, the matrix for L is simply the upper triangular matrix above. □

The converse is also true. If A is upper triangular, then you can easily check that $\operatorname{span}(\boldsymbol{e}_1, \dots, \boldsymbol{e}_k)$ is invariant under A for all k. So we see that a matrix can be triangularized if and only if there is an invariant filtration.

Lemma 2.2.7. *For any linear transformation $L : V \to V$ on a finite dimensional complex vector space V, there is an invariant filtration. (Note that this statement NEEDS V to be a complex vector space.)*

Proof. If $\dim V = 1$, this is trivial. We proceed by induction on $\dim V$.

Suppose $\dim V = n > 1$. Let $\boldsymbol{v}_1$ be any eigenvector for L for an eigenvalue λ_1. (Picking this $\boldsymbol{v}_1$ requires V to be a complex vector space, because some real matrices has no real eigenvectors.) Let V_1 be the subspace spanned by $\boldsymbol{v}_1$, and pick any complement subspace V_2 to V_1. Then L break downs into submaps $L_{ij} : V_j \to V_i$.

Consider $L_{22} : V_2 \to V_2$. Since $\dim V_2 = n-1$, by induction hypothesis, it has an L_{22}-invariant filtration, say $\{0\} = W_0 \subseteq W_1 \subseteq \cdots \subseteq W_{n-1} = \dim V_2$. I claim that $V_1 + W_k$ is L-invariant.

Obviously $L(V_1) \subseteq V_1 \subseteq V_1 + W_k$. So we only need to prove that $L(W_k) \subseteq V_1 + W_k$.

Pick any $\boldsymbol{w} \in W_k \subseteq V_2$, then L sends things in V_2 to V via L_{21} and L_{22}. So $L\boldsymbol{w} = L_{21}\boldsymbol{w} + L_{22}\boldsymbol{w}$. Now $L_{21}\boldsymbol{w} \in V_1$, while $L_{22}\boldsymbol{w} \in W_k$ because W_k is L_{22}-invariant. So $L\boldsymbol{w} \in V_1 + W_k$ indeed.

Now we can check that $\{\boldsymbol{0}\} \subseteq V_1 \subseteq V_1 + W_1 \subseteq \cdots \subseteq V_1 + W_{n-1} = V_1 + V_2 = V$ is the desired filtration. ☐

Note that, given any invariant filtration for A, simply let $\boldsymbol{v}_i$ be a unit vector orthogonal to V_{i-1} inside of V_i (like finding a normal vector to a plane in the space). Then we shall find a unitary matrix $B = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ such that $A = BTB^{-1}$ where T is upper triangular. This is the Schur decomposition theorem we did last semester. If you look into our proof last semester, you shall see that it is essentially IDENTICAl to what we are doing here.

2.2.2 Nilpotent Canonical Form

Definition 2.2.8. *A matrix J is an $d \times d$ Jordan block for the eigenvalue λ if* $J = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}_{d\times d}$.

(In the case where $\lambda = 0$, we also say it is a nilpotent Jordan block.)

Let us show that all nilpotent matrices can be block diagonalized where the diagonal blocks are nilpotent Jordan blocks.

Theorem 2.2.9. *If A is nilpotent, then we can find B such that $A = BDB^{-1}$ where D is block diagonal, and each diagonal block is a nilpotent Jordan block.*

Note that the nilpotent Jordan blocks are all "shift-up" operators, e.g., $\begin{bmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{bmatrix}$ would sends $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ to $\begin{bmatrix} y \\ z \\ 0 \end{bmatrix}$, it shifts the coordinates up. If we keep sending all the coordinates upwards, then eventually nothing will survive. In particular, if J is an $n \times n$ Jordan block, then it has a kill chain $\boldsymbol{e}_n \xmapsto{J} \dots \xmapsto{J} \boldsymbol{e}_1 \xmapsto{J} \boldsymbol{0}$ where the non-zero vectors for a basis.

In particular, our theorem says that any nilpotent matrix can be block diagonalized into such "shift-up" operators. Each Jordan block here has a corresponding "kill chain basis", and our matrix will have several kill chains whose non-zero vectors form a basis.

The next example here will show the algorithm to do the theorem above.

Example 2.2.10. Suppose A is a 7×7 nilpotent matrix. The chain of subspaces $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^2) \subseteq \mathrm{Ker}(A^3) \subseteq \mathrm{Ker}(A^4)$ has a chain of dimensions $3 \leq 5 \leq 6 \leq 7$. Note that this is NOT a filtration by itself, because some adjacent subspaces might differ by more than one dimensions.

Now we fill up the following chart from the bottum upwards:

$$\left(\begin{array}{c|ccc} \mathrm{Ker}(A)-\{\mathbf{0}\} & A^3v_1 & Av_2 & v_3 \\ \mathrm{Ker}(A^2)-\mathrm{Ker}(A) & A^2v_1 & v_2 & \\ \mathrm{Ker}(A^3)-\mathrm{Ker}(A^2) & Av_1 & & \\ \mathrm{Ker}(A^4)-\mathrm{Ker}(A^3) & v_1 & & \end{array}\right).$$

How did this work? We start by looking at the gap between $\mathrm{Ker}(A^4)$ and $\mathrm{Ker}(A^3)$. Note that the two subspace differ by exactly one dimension, so one extra vector is enough to extend $\mathrm{Ker}(A^3)$ to $\mathrm{Ker}(A^4)$. So we simply pick any $\boldsymbol{v}_1 \in \mathrm{Ker}(A^4) - \mathrm{Ker}(A^3)$.

Note that if $\boldsymbol{v}_1 \in \mathrm{Ker}(A^4) - \mathrm{Ker}(A^3)$, then we automatically have $A\boldsymbol{v}_1 \in \mathrm{Ker}(A^3) - \mathrm{Ker}(A^2)$, $A^2\boldsymbol{v}_1 \in \mathrm{Ker}(A^2) - \mathrm{Ker}(A)$ and $A^3\boldsymbol{v}_1 \in \mathrm{Ker}(A) - \{\mathbf{0}\}$. So we automatically filled a vector into each gap. We have $\mathrm{Ker}(A^4)$ spanned by $\mathrm{Ker}(A^3)$ and $\boldsymbol{v}_1$.

Now consider the gap between $\mathrm{Ker}(A^3)$ and $\mathrm{Ker}(A^2)$. Note that the two subspace differ by exactly one dimension, and we already have $A\boldsymbol{v}_1$ to fill in this gap, so there is nothing to do. We have $\mathrm{Ker}(A^3)$ spanned by $\mathrm{Ker}(A^2)$ and $A\boldsymbol{v}_1$.

Now consider the gap between $\mathrm{Ker}(A^2)$ and $\mathrm{Ker}(A)$. Note that the two subspace differ by two dimensions. We already have $A^2\boldsymbol{v}_1$ in this gap, but we need another vector. Pick any $\boldsymbol{v}_1 \in \mathrm{Ker}(A^2) - (\mathrm{Ker}(A) + \mathrm{span}(A^2\boldsymbol{v}_1))$. Now we have $\mathrm{Ker}(A^2)$ spanned by $\mathrm{Ker}(A)$ and $A^2\boldsymbol{v}_1, \boldsymbol{v}_2$.

Finally consider the gap between $\mathrm{Ker}(A)$ and $\{\mathbf{0}\}$. Note that the two subspaces differ by three dimensions. This time, we have $A^3\boldsymbol{v}_1, A\boldsymbol{v}_2$ in this gap already. I claim that they are linearly independent (proven in a later lemma), hence we just need one more. Pick any $\boldsymbol{v}_3 \in \mathrm{Ker}(A) - \mathrm{span}(A^3\boldsymbol{v}_1, A\boldsymbol{v}_2)$. Then we have $\mathrm{Ker}(A)$ spanned by $A^3\boldsymbol{v}_1, A\boldsymbol{v}_2, \boldsymbol{v}_3$.

Now, we see that the following subspaces are spanned by the following vectors:

$$\left(\begin{array}{c|ccccccc} \mathrm{Ker}(A) & A^3v_1 & Av_2 & v_3 & & & & \\ \mathrm{Ker}(A^2) & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 & & \\ \mathrm{Ker}(A^3) & Av_1 & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 & \\ \mathrm{Ker}(A^4) & v_1 & Av_1 & A^2v_1 & A^3v_1 & v_2 & Av_2 & v_3 \end{array}\right).$$

And furthermore, we have kill chains $\boldsymbol{v}_1 \overset{A}{\longmapsto} A\boldsymbol{v}_1 \overset{A}{\longmapsto} A^2\boldsymbol{v}_1 \overset{A}{\longmapsto} A^3\boldsymbol{v}_1 \overset{A}{\longmapsto} \mathbf{0}$, and $\boldsymbol{v}_2 \overset{A}{\longmapsto} A\boldsymbol{v}_2 \overset{A}{\longmapsto} \mathbf{0}$, and finally $\boldsymbol{v}_3 \overset{A}{\longmapsto} \mathbf{0}$. All the vectors in these three kill chains (other than the zero vectors) are linearly independent, and all the important invariant subspaces are spanned by these vectors in very nice manners.

Pick a basis $A^3\boldsymbol{v}_1, A^2\boldsymbol{v}_1, A\boldsymbol{v}_1, \boldsymbol{v}_1, A\boldsymbol{v}_2, \boldsymbol{v}_2, \boldsymbol{v}_3$, then you can check yourself that our matrix A would change into the following:

$$\left[\begin{array}{cccc|cc|c} 0 & 1 & & & & & \\ & 0 & 1 & & & & \\ & & 0 & 1 & & & \\ & & & 0 & & & \\ \hline & & & & 0 & 1 & \\ & & & & & 0 & \\ \hline & & & & & & 0 \end{array}\right].$$

☺

Two things could go wrong here. First of all, when we fill in the gap between $\mathrm{Ker}(A)$ and $\{\mathbf{0}\}$, we need $A^3\boldsymbol{v}_1$ and $A\boldsymbol{v}_2$ to be linearly independent. Why is that?

Recall that we picked $\boldsymbol{v}_2$ such that $\mathrm{Ker}(A), A^2\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent. It turns out that this is enough.

Lemma 2.2.11. *If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \mathrm{Ker}(A^t)$ are linearly independent, then $A\boldsymbol{v}_1, \ldots, A\boldsymbol{v}_k, \mathrm{Ker}(A^{t-1})$ are linearly independent.*

Proof. Suppose $(\sum a_i A\boldsymbol{v}_i) + b\boldsymbol{w} = \mathbf{0}$ where $\boldsymbol{w} \in \mathrm{Ker}(A^{t-1})$. Apply A^{t-1} on both sides. Then we have $(\sum a_i A^t\boldsymbol{v}_i) + bA^{t-1}\boldsymbol{w} = \mathbf{0}$, and here $A^{t-1}\boldsymbol{w}$ would die.

So we have $A^t(\sum a_i \boldsymbol{v}_i) = \mathbf{0}$. This implies that $\sum a_i \boldsymbol{v}_i = \boldsymbol{w}'$ for some $\boldsymbol{w}' \in \mathrm{Ker}(A^t)$. But since these $\boldsymbol{v}_i$ and $\mathrm{Ker}(A^t)$ are linearly independent, this means all $a_i = 0$ and $\boldsymbol{w}' = \mathbf{0}$.

This in turn means that, from the equation $(\sum a_i A\boldsymbol{v}_i) + b\boldsymbol{w} = \mathbf{0}$, we must have $b\boldsymbol{w} = \mathbf{0}$. So if $\boldsymbol{w}$ is non-zero, $A\boldsymbol{v}_1, \ldots, A\boldsymbol{v}_k, \boldsymbol{w}$ are linearly independent. $\square$

This lemma guarantees that our algorithm in the example shall always work, and hence our theorem is correct.

2.3 Jordan Canonical Form

The Jordan canonical form simply combines all previous results. There is one last simple lemma.

Lemma 2.3.1. *If all eigenvalues of A are λ, then $A = BJB^{-1}$ where J is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ.*

Proof. All eigenvalues of $A-\lambda I$ are zero, so this is nilpotent. So $A-\lambda I = BJB^{-1}$ where J is block diagonal, and each diagonal block is a nilpotent Jordan block. Then $A = BJB^{-1} + \lambda I = B(J+\lambda I)B^{-1}$. And we can see that $J+\lambda I$ is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ. $\square$

Theorem 2.3.2 (Jordan canonical form)**.** *For any matrix A, we have $A = BJB^{-1}$ where J is block diagonal, and each diagonal block of J is a Jordan block.*

Proof. Since the domain is the direct sum of generalized eigenspaces, we can assume that $A = XDX^{-1}$ where $D = \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix}$ is block diagonal, and each diagonal block D_i corresponds to a generalized eigenspace for the eigenvalue λ_i.

So all eigenvalues of D_i are λ_i. So $D_i = B_i J_i B_i^{-1}$ where J_i is block diagonal, and each diagonal block is a Jordan block with eigenvalue λ_i.

Then $A = BJB^{-1}$ where $B = X \begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_k \end{bmatrix}$ and $J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{bmatrix}$ is block diagonal, and each diagonal block of J is a Jordan block. $\square$

How to find Jordan canonical form? Let us have some calculation examples.

Lemma 2.3.3. *If λ is an eigenvalue of A with algebraic multiplicity m, then $N_\infty(A-\lambda I) = \mathrm{Ker}(A-\lambda I)^m$.*

Proof. Take $N_\infty(A-\lambda I)$ as the domain, and consider the operator $A-\lambda I$, which is nilpotent. So if $\dim N_\infty(A-\lambda I) = m$, then $(A-\lambda I)^m = 0$ on the space $N_\infty(A-\lambda I)$.

So now using our original domain, we see that $N_\infty(A-\lambda I) \subseteq \mathrm{Ker}(A-\lambda I)^m$. But by definition $\mathrm{Ker}(A-\lambda I)^m \subseteq N_\infty(A-\lambda I)$. So we are done. $\square$

Example 2.3.4. Consider $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$. Then $\det(xI-A) = \det \begin{bmatrix} x-2 & 0 & 0 \\ 1 & x-1 & -2 \\ -3 & 0 & x-1 \end{bmatrix} = (x-2)\det \begin{bmatrix} x-1 & -2 \\ 0 & x-1 \end{bmatrix} = (x-2)(x-1)^2$. So it has eigenvalue 1 with algebraic multiplicity 2 and eigenvalue 2 with algebraic multiplicity 1. So it must has a generalized eigenspace V_1 for the eigenvalue 1 of dimension 2 and a generalized eigenspace V_2 for the eigenvalue 2 of dimension 1.

What is V_1? It is $\mathrm{Ker}(A-I)^2 = \mathrm{Ker} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 2 \\ 3 & 0 & 0 \end{bmatrix}^2 = \mathrm{Ker} \begin{bmatrix} 1 & 0 & 0 \\ 5 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}$, which is spanned by $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. And restricted to this subspace V_1, under this basis, we have $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$, and $A - I = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$ in indeed nilpotent. Now our theorem on nilpotent Jordan normal form tells us that we could pick basis $\boldsymbol{v}_1 = (A-I)\boldsymbol{v}_2$ and $\boldsymbol{v}_2 = \begin{bmatrix}0\\0\\1\end{bmatrix}$ as the right basis for V_1.

What is V_2? It is $\operatorname{Ker}(A-2I) = \operatorname{Ker}\begin{bmatrix}0&0&0\\-1&-1&2\\3&0&-1\end{bmatrix}$ which is spanned by $\boldsymbol{v}_3 = \begin{bmatrix}1\\5\\3\end{bmatrix}$. Obviously A restricted to V_2 is just $\begin{bmatrix}2\end{bmatrix}$ and there is nothing to do here.

So the best basis for V should be $(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3) = \begin{bmatrix}0&0&1\\2&0&5\\0&1&3\end{bmatrix}$. And under this basis, the new matrix for A should be $\begin{bmatrix}1&1&0\\0&1&0\\0&0&2\end{bmatrix}$. Let us check this. Indeed, we have:

$$\begin{bmatrix}0&0&1\\2&0&5\\0&1&3\end{bmatrix}\begin{bmatrix}1&1&0\\0&1&0\\0&0&2\end{bmatrix}\begin{bmatrix}0&0&1\\2&0&5\\0&1&3\end{bmatrix}^{-1} = \begin{bmatrix}0&0&2\\2&2&10\\0&1&6\end{bmatrix}\begin{bmatrix}-5/2&1/2&0\\-3&0&1\\1&0&0\end{bmatrix} = \begin{bmatrix}2&0&0\\-1&1&2\\3&0&1\end{bmatrix} = A.$$

☺

Example 2.3.5. Let us have a more complicated example. Feel free to have a matrix calculator in hand while reading this example.

Say $A = \begin{bmatrix}-10&9&-7&-1&7\\-17&13&-9&-2&12\\-14&9&-6&-1&10\\-13&9&-7&0&9\\-12&9&-7&-1&9\end{bmatrix}$. You can find its characteristic polynomial and check that its eigenvalues are 1,1,1,1,2.

The eigenvalue 2 is simple. It has algebraic and geometric multiplicity 1, and you can find its corresponding eigenvector is $\boldsymbol{v}_5 = \begin{bmatrix}5\\9\\8\\7\\6\end{bmatrix}$.

For the eigenvalue 1, consider $A-I = \begin{bmatrix}-11&9&-7&-1&7\\-17&12&-9&-2&12\\-14&9&-7&-1&10\\-13&9&-7&-1&9\\-12&9&-7&-1&8\end{bmatrix}$. You can check that $\dim\operatorname{Ker}(A-I) = 2, \dim\operatorname{Ker}(A-I)^2 = 3, \dim\operatorname{Ker}(A-I)^3 = 4$, and we don't need to continue once we reach dimension 4, because 1 only has algebraic multiplicity 4.

Pick any $\boldsymbol{v}_3 \in \operatorname{Ker}(A-I)^3 - \operatorname{Ker}(A-I)^2$, and set $\boldsymbol{v}_2 = (A-I)\boldsymbol{v}_3$ and $\boldsymbol{v}_1 = (A-I)\boldsymbol{v}_2$, and find any $\boldsymbol{v}_4$ such that $\boldsymbol{v}_1, \boldsymbol{v}_4$ span $\operatorname{Ker}(A-I)$. One possible choice is $\boldsymbol{v}_3 = \begin{bmatrix}1\\2\\2\\1\\1\end{bmatrix}$, then $\boldsymbol{v}_2 = \begin{bmatrix}-1\\-1\\-1\\-1\\-1\end{bmatrix}$, and then $\boldsymbol{v}_1 = \begin{bmatrix}3\\4\\3\\3\\3\end{bmatrix}$.

Then you can pick say $\boldsymbol{v}_4 = \begin{bmatrix}1\\3\\3\\2\\1\end{bmatrix}$. Note that since $\boldsymbol{v}_3$ goes to $\boldsymbol{v}_2$, which goes to $\boldsymbol{v}_1$, and $\boldsymbol{v}_4$ stands alone, therefore the corresponding nilpotent Jordan block is $\left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right]$.

So under the basis $B = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5) = \begin{bmatrix} 3 & -1 & 1 & 1 & 5 \\ 4 & -1 & 2 & 3 & 9 \\ 3 & -1 & 2 & 3 & 8 \\ 3 & -1 & 1 & 2 & 7 \\ 3 & -1 & 1 & 1 & 6 \end{bmatrix}$, we have A in Jordan canonical form $J = \left[\begin{array}{ccc|c|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 2 \end{array}\right]$. In particular, $A = BJB^{-1}$.

(Obviously I designed it so that we could have integer solutions.... Usually we should not be so lucky. A super interesting challenge question here: Can you design another 5×5 integer-entry matrix A such that for $A = BJB^{-1}$, both B and J has integer entries?) ☺

Sometimes we can tell the Jordan normal form right away, just from the geometric and algebraic multiplicity. Here is why.

Proposition 2.3.6. *Suppose λ is an eigenvalue of A with algebraic mulitiplicity m_a and geometric multiplicity m_g. Then m_g is the number of λ-Jordan blocks in the Jordan canonical form of A, while m_a is the sum of the sizes of all these Jordan blocks.*

Proof. WLOG we can assume that A is already in Jordan canonical form. Furthermore, all the blocks not related to λ are irrelavant. It is then clear that each λ-Jordan block contributes to exactly one dimension to $\mathrm{Ker}(A - \lambda I)$, so the statement about geometric multiplicity is done.

The statement about algebraic multiplicity is trivial by just looking at the characteristic polynomial of block diagonal matrices. □

In particular, if a matrix A has all geometric multiplicities equal to algebraic multiplicities, then the number of λ-blocks would equal to the sum of sizes of all these blocks, i.e., each block is 1×1. So the matrix is diagonalizable.

Example 2.3.7. For example, consider $J = \begin{bmatrix} 1 & 1 & & \\ 0 & 1 & & \\ & & 1 & \\ & & & 2 \end{bmatrix}$. Chech that the eigenvalue 1 here has indeed geometric multiplicity 2 and algebraic multiplicity 3.

Conversely, suppose A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 3, m_g(2) = m_a(2) = 1$, then it has two 1-blocks and a single 2-block. Furthermore, since the two 1-blocks have a total size 3, it must be $1 + 2$. So A must have the Jordan canonical form J above.

Of course, if A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 4, m_g(2) = m_a(2) = 1$, then there is no way to tell now. The two 1-blocks could be $2 + 2$ or $1 + 3$, and we may never know. ☺

The last example where we cannot decide is unfortunate. However, there is one more tool we can use: the minimal polynomial. But before we do that, let us do the famous Cayley-Hamilton Theorem as a corollary to our study of generalized eigenspaces.

Corollary 2.3.8 (Cayley-Hamilton Theorem)**.** *For any matrix A, let $p_A(x)$ be its characteristic polynomial. Then $p_A(A)$ is the zero matrix.*

Proof. It is enough to show that $p_A(A)$ kills each generalized eigenspace.

For each eigenvalue λ, let m be its algebraid multiplicity. Then $p_A(x) = q(x)(x-\lambda)^m$. So $p_A(A) = q(A)(A-\lambda I)^m$.

So if $\boldsymbol{v} \in N_\infty(A-\lambda I) = \mathrm{Ker}(A-\lambda I)^m$, then $p_A(A)\boldsymbol{v} = q(A)(A-\lambda I)^m\boldsymbol{v} = \mathbf{0}$.

But this is true for all λ. So $p_A(A)$ kills all vectors in all generalized eigenspaces. Oops. $\square$

Definition 2.3.9. *We say a polynomial $p(x)$ is a killing polynomial for A if $p(A) = 0$. We say $p(x)$ is a minimal polynomial for A if any killing polynomial of A must contain $p(x)$ as a factor.*

Proposition 2.3.10. *Any square matrix A has a minimal polynomial.*

Proof. Suppose that A is in Jordan normal form. Then $p(A)$ is simply applying $p(x)$ to each diagonal block, and $p(x)$ is a killing polynomial if and only if it kills all blocks simultaneously. So it is enough to prove this statement for each Jordan block.

Suppose A be a single Jordan block, say $n \times n$ with eigenvalue λ. Then $A - \lambda I$ is the shift up operator, and if $p(x) = a_0 + a_1x + \cdots + a_kx^k$, then $p(A-\lambda I)$ will have diagonal entries a_0, and entries right above the diagonal a_1, and so on so forth. So $p(A-\lambda I) = 0$ if and only if the coefficients $a_0, \ldots, a_{n-1}$ are zero, i.e., $p(x)$ contains x^n as a factor. So if $q(A) = 0$, then $q(A)$ must contains $(A-\lambda I)^n$ as a factor.

To sum up, to kill a Jordan block, say $n \times n$ with eigenvalue λ, $p(x)$ must contain factor $(x-\lambda)^n$.

So $p(x)$ kills A if and only if it contains $(x-\lambda)^{m_\lambda}$ for all λ, where λ is the size of largest λ-Jordan block for A. $\square$

Example 2.3.11. If A is any matrix with eigenvalue $1, 2$, and $m_g(1) = 2, m_a(1) = 4, m_g(2) = m_a(2) = 1$, then there is no way to tell now. The two 1-blocks could be $2+2$ or $1+3$, and we may never know.

But if we also know that the minimal polynomial is $(x-1)^2(x-2)$, then the 1-blocks must be $2+2$, and we must have two 1-blocks of size 2, and a single 2-block of size 1. If the minimal polynomial is $(x-1)^3(x-2)$, then the 1-blocks must be $1+3$, and we must have a 1-blocks of size 3, a 1-blocks of size 1, and a single 2-block of size 1.

Of course, there will be situations where even the minimal polynomial is not enough. Suppose A is 7×7 with $m_a(1) = 7, m_g(1) = 3$, and minimal polynomial $(x-1)^3$. Then it could be $3+3+1$ or $3+2+2$, and we cannot tell anymore. Time to get your hand dirty and actually compute those blasted $\mathrm{Ker}(A-I)^k$. ☺

2.4 (Optional) The geometric interpretation of Jordan canonical form and generalized eigenspaces

Technically we are done. The theorem of Jordan canonical form is saying that, for any linear map, we can decompose it into independent "submaps" that are Jordan blocks. So if we understand all Jordan blocks we would understand every single matrix.

So this raises a new question. How would a Jordan block behave? Let us look at a few to generate some ideas.

Example 2.4.1. What are nilpotent Jordan blocks? Consider the 3×3 nilpotent Jordan block N. It sends the z-axis to the y-axis, and the y-axis to the x-axis. Huh, it seems to be rotating. But then it sends the x-axis to zero. So we are "rotating inwards to zero". (Nei Juan....)

Personally I think of $\mathbb{R}^3$ as the space of all students, and N as some competitive and selective process. Then after N, all students are squeezed into the xy-plane, trying to excel. After another N, now everyone is squeezed into the x-axis, trying to be the best of the best. After yet another N, everyone dies of exhaustion apparently.... ☺

Example 2.4.2. $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is the standard shearing. In general, consider $E = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$. It sends rectangles, with sides parallel to the coordinate-lines, into parallelograms of the same height. Draw a few graphic examples and shapes to see this better. This process would preserve the base and height of the parallelogram, so it preserves the area.

(Also note that EA is a row operation on A. Such row operations corresponds to shearings, so it preserves area, and hence it preserves the determinant. I.e., $\det(EA) = \det(A)$.)

If you repeatedly apply $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ to a vector, say $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, you get $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}$, and so on. Basically the second coordinates are always the same, while the first coordinate keep progressing. The so the orbits of A are lines parallel to the x-axis. ☺

Example 2.4.3. Now consider $J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. It sends $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, then to $\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$, then to $\begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix}$, then to $\begin{bmatrix} 6 \\ 4 \\ 1 \end{bmatrix}$, and so on. This is EXACTLY the left three entries of the Pascal's triangule (Yang Hui triangle, or binomial coefficients, etc.)!

So to see $J^k \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, you can imagine that you are doing $(x+1)^k$, and read out the last three coefficients.

You can also see that $J^k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} k \\ 1 \\ 0 \end{bmatrix}^T$, which is basically the last three coefficients of $x(x+1)^k$. In general, $J^k \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is the last three coefficients of $(ax^2+bx+c)(x+1)^k$. Funny, no?

Is this really true? Well, let P_2 be set of polynomials mod x^3. I.e., we consider two polynomials to be the same as long as they have the same coefficients at degree $2, 1, 0$. For example, we think of $x^3 + x + 1$ and $x^4 + x + 1$ as the same element in P_2.

Then clearly P_2 is three dimensional, hence we can idenfity it with $\mathbb{R}^3$ via its standard basis $x^2, x, 1$. Then how does J behaves on P_2? It sends 1 to $x + 1$, and x to $x^2 + x$, and x^2 to x^2, which is the same as $x^3 + x^2$ since we only care about the coefficients at degree $2, 1, 0$. So J behaves exactly by multiplying polynomials by $(x + 1)$. So $J^k(ax^2 + bx + c) = (ax^2 + bx + c)(x + 1)^k \pmod x)^3$.

This algebraic picture can be generalized to Jordan blocks with eigenvalue 1 of arbitrary size. ☺

Example 2.4.4. What is the geometric behavior of $J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$? Say what are its orbits (smooth curves C such that J always maps each point in C back to some point in C)?

Well, in general, $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ would goes to $\begin{bmatrix} a+b \\ b+c \\ c \end{bmatrix}$, and then to $\begin{bmatrix} a+b+b+c \\ b+c+c \\ c \end{bmatrix}$, and then to $\begin{bmatrix} a+b+(b+c)+(b+c+c) \\ b+c+c+c \\ c \end{bmatrix}$ and so on. So after k steps, J^k would maps it to $\begin{bmatrix} a+kb+(0+1+\cdots+(k-1))c \\ b+kc \\ c \end{bmatrix} = \begin{bmatrix} a+kb+\frac{1}{2}(k^2-k)c \\ b+kc \\ c \end{bmatrix}$.

So generically, to find orbits, I simply replace the integer k by an arbitrary real number t, and we have the orbits $p(t) = \begin{bmatrix} \frac{c}{2}t^2 + (b-\frac{c}{2})t + a \\ ct+b \\ c \end{bmatrix}$. It is easy to verify that any points on this curve shall stay on this curve after J.

As you can see, the third coordinate never change, so the orbit curves stays on a plane (parallel to the xy-plane). On this plane, the first coordintae is in fact a degree two polynomial of the second coordinate. So on this plane, we would actually see a graph of a parabola. So orbits of J are various parabolas parallel to the xy-plane.

Note that for each parabola on a plane $z = c \neq 0$, when $t = -\frac{b}{c}$, then the parabola would go through the

xz-plane. So if you want to find all parabolas on the plane $z = c$, then they are $p(t) = \begin{bmatrix} \frac{c}{2}t^2 - \frac{c}{2}t + a \\ ct \\ c \end{bmatrix}$, or the parabola $p(t) = \begin{bmatrix} \frac{c}{2}t^2 - \frac{c}{2}t \\ ct \\ c \end{bmatrix}$ shifted along the x-axis. Furthermore, since we only care about the curve, not how it is parametrized, we can further more substitute t by t/c. Then we have $p(t) = \begin{bmatrix} \frac{1}{2c}t^2 - \frac{1}{2}t \\ t \\ c \end{bmatrix}$ shifted along the x-axis.

So for each constant c, the orbits on $z = c$ are just parabolas obtained by translating this along the x-axis.

I highly recommand you to draw these parabolas on $z = 1, z = 2, z = -1$ to see what would happen. Also fell free to draw the picture on the plane $z = 0$, and see why this is the limiting case for $z > 0$ and $z < 0$.

If you want to see the geometric behavior, you can try to generlize this further. Say you want a size 4 Jordan block with eigenvalue 1. Then for any orbit curve, again the last coordinate is constant for some $d \in \mathbb{C}$. If the third coordinate is t, then the second coordinate would again be $\frac{1}{2d}t^2 - \frac{1}{2}t$ shifted around by some constant. And finally, the first coordinate would be a degree 3 polynomial in t. It would look like some form of spiral. Consider curves like $\begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} \in \mathbb{R}^4$ for a idea of this kind of spirals. ☺

Example 2.4.5. Consider a Jordan block with eigenvalue, say $J = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$. Then it sends $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$, then to $\begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$ and so on. It looks like you are doing $(x+2)^k$.

Indeed, algebraically $J^k \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is the last three coordinates of $(ax^2 + bx + c)(x+2)^k$ for the same reason as before. Now you can generalize this to get the algebraic behavior of all Jordan blocks of all size for all eigenvalues.

What about its geometric behavior? Suppose we start at some vector $\begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix}$, and we construct $J \begin{bmatrix} a_{n-1} \\ b_{n-1} \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix}$. Then we see that $c_n = 2^n c_0$.

We can see that $b_n = 2b_{n-1} + c_{n-1}$. Divide this by 2^n on both sides (because we know all three sequences must be related to 2^n somehow, as 2 is the eigenvalue), we see that $\frac{b_n}{2^n} = \frac{b_{n-1}}{2^{n-1}} + \frac{c_0}{2}$. So the sequence $\frac{b_n}{2^n}$ is arithmetic and $\frac{b_n}{2^n} = \frac{b_0}{2^0} + \frac{c_0}{2}n$. So $b_n = 2^n b_0 + n2^{n-1}c_0$.

Finally, $a_n = 2a_{n-1} + b_{n-1}$. By a similar argument, $\frac{a_n}{2^n} = \frac{a_{n-1}}{2^{n-1}} + \frac{b_0}{2} + (n-1)\frac{c_0}{4}$. So $\frac{a_n}{2^n}$ is a degree two polynomial in n, and specifically you can see that $\frac{a_n}{2^n} = \frac{b_0}{2}n + \frac{c_0}{4}(0+1+2+...+(n-1)) = \frac{c_0}{8}n^2 + (\frac{b_0}{2} - \frac{c_0}{8})n$. So $a_n = n^2 2^{n-3}c_0 + n2^{n-3}(4b_0 - c_0)$.

So a typical curve looks like $p(t) = \begin{bmatrix} t^2 2^{t-3}c_0 + t2^{t-3}(4b_0 - c_0) \\ 2^t b_0 + t2^{t-1}c_0 \\ 2^t c_0 \end{bmatrix}$. By a change in parametrization, we can choose $2^t c_0$ as the new parameter t, then the curve is $p(t) = t\begin{bmatrix}a(t)\\b(t)\\c(t)\end{bmatrix}$ where $a(t), b(t), c(t)$ here are polynomials in $\ln t$ of degree 2,1,0.

Also note that, asymptotically for super large n, $\lim \frac{b_n^2}{2a_n c_n} = 1$. Therefore these curves has asymptotic surface $xz = y^2$. What is this surface? It is a cone around the line $\{y = 0\} \cap \{x = z\}$. So all these orbital curves will eventually get closer and closer to this cone. ☺

Example 2.4.6. As shown in the example above, the geometric picture of a Jordan block is not always easy to compute. However, let us try to do another case, $J = \begin{bmatrix}\lambda & 1 & 0\\0 & \lambda & 1\\0 & 0 & \lambda\end{bmatrix}$ for some extremely large λ. Then since λ is so large, comparatively the ones are ignorable. So $J \approx \lambda I$. This geometric picture is very easy now, it is approximately just stretch everything by λ. So the orbits are approximately just rays shooting from the origin, with some minor perturbations. ☺

The process of finding Jordan canonical form is equivalent to this: First we find generalized eigenspaces of A. Next, for each generalized eigenspace for an eigenvalue λ, we identify linearly independent killing chains of $A - \lambda I$.

With this in mind, what is the generalized eigenspace, i.e., vectors eventually killed by $A - \lambda I$? Here let us formulate an alternative definition for generalized eigenspaces.

The most fundamental motivation for studying eigenstuff is to understand the behavior of sequences like $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$.

Example 2.4.7. Again consider $A = \begin{bmatrix}1 & 1 & 0\\0 & 1 & 1\\0 & 0 & 1\end{bmatrix}$. We know that its orbits are parabolas. In particular, the sequence $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$ would tend to produce longer and longer vectors, and they would never converge.

However, even though the vectors do not converge, their DIRECTIONS would in fact converge! The directions of these vectors would get closer and closer to the direction of the opening for the parabola, which is always in the direction of plus or minus x-axis.

In particular, the directions of $\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$ converge to $\pm\boldsymbol{e}_1$. ☺

Definition 2.4.8. *Given an inner product space (say, $\mathbb{C}^n$ with the dot product if you prefer), and any linear transformation A, and any vector $\boldsymbol{v}$, we set $\boldsymbol{v}_0 = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$, and set $\boldsymbol{v}_{i+1} = \frac{A\boldsymbol{v}_i}{\|A\boldsymbol{v}_i\|}$. Then if the limit exists and $\lim_{t\to\infty} \boldsymbol{v}_t = \boldsymbol{w}$, then we say $\boldsymbol{v}$ converges in direction to $\boldsymbol{w}$ under iterations of A.*

Proposition 2.4.9. *If $\boldsymbol{v}$ is in a generalized eigenspace for some eigenvalue $\lambda > 0$ of A, then it converges in direction to some eigenvector of λ under iterations of A.*

This can be proven easily with basic topology, which is outside of the scope of this class. (The unit sphere is compact, the rest is easy.) Of course we cannot do that here. So now let us prove this using linear algebra instead.

Proof. Suppose $\boldsymbol{v}$ is in the generalized eigenspace for the eigenvalue λ. Then $A - \lambda I$ would kill it in finitely many steps, say $\boldsymbol{v} \mapsto (A - \lambda I)\boldsymbol{v} \mapsto \cdots \mapsto (A - \lambda I)^{k-1}\boldsymbol{v} \mapsto \boldsymbol{0}$ where $(A - \lambda I)^{k-1}\boldsymbol{v} \neq \boldsymbol{0}$.

Let V be the span of $\boldsymbol{v}, (A-\lambda I)\boldsymbol{v}, \ldots, (A-\lambda I)^{k-1}\boldsymbol{v}$. (Recall that these vectors are linearly independent.) It should be very obvious that V is a k-dimensional $(A-\lambda I)$-invariant subspace. Hence it is also A-invariant. Furthermore, if we restrict the domain and codomain to V, and use basis $(A - \lambda I)^{k-1}\boldsymbol{v}, \ldots, \boldsymbol{v}$, then $A - \lambda I$ would have matrix $\begin{bmatrix}0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & 0\end{bmatrix}$. As a result, A would have a matrix of $\begin{bmatrix}\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda\end{bmatrix}$. This is a $k \times k$ matrix.

So our problem is reduced to this: we can assume that the space we study is $\mathbb{C}^k$, and the matrix is simple a single Jordan block $\begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$. Set $\boldsymbol{v} = \boldsymbol{e}_k$ the last standard basis vector, we aim to show that the sequence $\boldsymbol{v}_i$ as defined would converge to an eigenvector. Also note that the only eigenvectors in this space are multiples of $\boldsymbol{e}_1$.

Note that $\boldsymbol{v}_t$ is in the same direction of $A^t\boldsymbol{v}$, and its coordinates should corresponds to the last n coefficients of the polynomial $(x+\lambda)^t$. By the binomial theorem, it is $\begin{bmatrix} \binom{t}{k-1}\lambda^{t-k+1} \\ \vdots \\ \binom{t}{0}\lambda^t \end{bmatrix}$.

(Notation: Here $\binom{n}{k}$ means the number of ways to choose k objects out of n objects. In Chinese textbooks it is more traditionally written as C_n^k, the combinatorial number to choose k out of n.)

Since we only care about the direction, we can divide all coordinates by the same constant λ^t. Then we are looking at the vector $\begin{bmatrix} \binom{t}{k-1}\lambda^{-k+1} \\ \vdots \\ \binom{t}{0}\lambda^0 \end{bmatrix}$. Note that the coordinates are all polynomials of t, and the i-th coordinate is a polynomial of t of degree $k-i$. In particular, as $t \to \infty$, eventually the first coordinate (polynomial of degree $k-1$) will outgrow everyone else (polynomials of lower degree). So the direction converge towards $\boldsymbol{e}_1$ indeed. $\square$

The proof above should clarify the following idea: generalized eigenspace is where killing chains happen (for the corresponding $A - \lambda I$). And each killing chain corresponds to some indecomposable invariant subspace (cannot be the direct sum of two smaller invariant subspaces), on which the linear map will be a Jordan block. In this sense, Jordan blocks are indeed the "atoms" of a linear map.

What if $\lambda = 0$? Then the sequence $A^t\boldsymbol{v}$ is going to be $\mathbf{0}$ in finitely many steps. So it does not converge to any direction, since it becomes zero.

What if $\lambda < 0$? By basically the same proof the sequence $(A^t\boldsymbol{v})$ for all even t is going to converge to an "eigendirection", while for all odd t the sequence will converge to the negation of the previous direction. It is "alternating", but they all converge to the same "eigenline".

Proposition 2.4.10. *Again suppose we have an inner product space, say $\mathbb{C}^n$ with dot product.*

Suppose V is an A-invariant subspace of $\mathbb{C}^n$ in which all non-zero vectors converge in direction to some eigenvector of $\lambda > 0$, then V is inside the generalized eigenspace of λ for A.

(In short, the generalized eigenspace of $\lambda > 0$ is the UNIQUE LARGEST A-invariant subspace, where all vectors converges in direction to some λ-eigendirection.)

Proof. Pick any $\boldsymbol{v} \in V$. Then since it is A-invariant, linear combinations of $\boldsymbol{v}, A\boldsymbol{v}, \dots$ are all in V. In particular, $(A-\lambda I)^n\boldsymbol{v} \in V$. Suppose $(A-\lambda I)^n\boldsymbol{v}$ is non-zero, then it converges in direction to an eigenvector of λ.

Now note that the whole domain decomposes as a direct sum of $N_\infty(A-\lambda I)$ and $R_\infty(A-\lambda I)$. Then we have a corresponding decomposition $\boldsymbol{v} = \boldsymbol{v}_N + \boldsymbol{v}_R$. Then $(A-\lambda I)^n\boldsymbol{v} = (A-\lambda I)^n\boldsymbol{v}_N + (A-\lambda I)^n\boldsymbol{v}_R = (A-\lambda I)^n\boldsymbol{v}_R$. Since $R_(A-\lambda I)$ is A-invariant, we would still have $(A-\lambda I)^n\boldsymbol{v}_R \in R_\infty(A-\lambda I)$. As a result, we have $(A-\lambda I)^n\boldsymbol{v} \in R_\infty(A-\lambda I)$.

In particular, if $(A-\lambda I)^n\boldsymbol{v}$ converges in direction to some unit vector, that unit vector must still be inside $R_\infty(A-\lambda I)$. But since it is also in V, it must converge in direction to some unit vector in $N_\infty(A-\lambda I)$. Contradiction.

Hence we must conclude that $(A-\lambda I)^n\boldsymbol{v} = \mathbf{0}$, which means $\boldsymbol{v} \in N_\infty(A-\lambda I)$. $\square$

The other cases are similar. We put the result here without proof.

If $\lambda = 0$, then the generalized eigenspace is the UNIQUE LARGEST A-invariant subspace on which A eventually kills everything.

And if $\lambda < 0$, then the generalized eigenspace is the UNIQUE LARGEST A-invariant subspace, where all vectors converges alternatingly to some λ-eigenline.

So we have a geometric description of generalized eigenspaces.

Example 2.4.11. The requirement that A-invariance is important! There are indeed (non-invariant) subspaces OUTSIDE of the generalized eigenspace for λ, where all non-zero vectors converges to λ-eigendirections.

Consider $A = \begin{bmatrix} 1 & 1 & \\ 0 & 1 & \\ & & 2 \end{bmatrix}$. I claim that all vectors NOT in the xy-plane would converge in direction to $\boldsymbol{e}_3$.

To see this, say we started with $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ where $c \neq 0$. Then $A^t\boldsymbol{v} = \begin{bmatrix} a+tb \\ b \\ c2^t \end{bmatrix}$. CLearly the last coorinate would dominate, and the vector's direction would be closer and closer to $\boldsymbol{e}_3$.

In general, if $\boldsymbol{v}$ has non-zero components involving many different generalized eigenspaces, then the λ with largest absolute value would dominate the convergence behavior of $\boldsymbol{v}, A\boldsymbol{v}, \ldots$.

What if some eigenvalues involved with $\boldsymbol{v}$ have the same absolute value? Then something funny might happen. Consider $\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$. Then other than the eigenvectors, nothing else would converge to eigendirections or even eigenlines. They simply bounce. ☺

2.5 (Optional) A naive QR-algorithm: How to find eigenvalues

You have been lied to. Your teacher might taught you this: to find eigenvalues of a matrix, first find its characteristic polynomial, and then find the roots.

But wait, what about the following famous theorem?

Theorem 2.5.1 (Abel-Ruffini)**.** *There is no algebraic solution to polynomials of degree 5 and above. I.e., using addition, substraction, multiplication, division, and k-th roots, it is impossible to have a formula for solving a generic polynomials of degree 5 and above.*

Remark 2.5.2. *You can think of this theorem as stating that traditional algebraic calculations, such as addition, substraction, multiplication, division, and k-th roots, are NOT expressive enough. For the polynomial $x^5 - x - 1$, its roots CANNOT be expressed. Therefore no generic formula exists if we are only allowed to use these calculations.*

What if we use other calculations? Well, they usually results in circular logic. For example, the Bring radical might help express the root. But how is the Bring radical calculated? It is calculated by solving polynomials of degree 5. Oops, circluar. All such attempts usually end up just defining the answer using the answer itself. They will NOT be helpful.

So if you have an $n \times n$ matrix where $n \geq 5$, you cannot really solve the characteristic polynomial. There is no such formula!

One might be tempted to approximate the roots of a high degree polynomial using a computer. But do you know how a computer approximate roots?

Definition 2.5.3. *Given a polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$, we define its companion matrix as $C_p = \begin{bmatrix} 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & \ldots & 0 & -c_1 \\ 0 & 1 & \ldots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -c_{n-1} \end{bmatrix}$. (Sometimes some people define the companion matrix as C_p^{T} instead. It rarely matters.)*

The idea is that C_p has characteristic polynomial $p(x)$ (see if you can prove this). Then the computer would actually tries to approximate eigenvalues of C_p, which are roots of $p(x)$. It is the other way around!

Finding roots cannot help you with eigenvalues. It is the eigenvalues that help you find roots.

So how can we find eigenvalues? The easy idea is the structure we mentioned above. If we pick an arbitrary $\boldsymbol{v}$, and compute $A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots$, then the direction would most likely converge to a eigenvector of the largest (in absolute value) eigenvalue. Since the sequence does not actually converge, only the direction converges, therefore we are going to normalize at each step. Simply put, we do the following:

Example 2.5.4. Pick a randome vector $\boldsymbol{v}_0 = \boldsymbol{v}$, and let $\boldsymbol{v}_{k+1} = \frac{A\boldsymbol{v}_k}{\|A\boldsymbol{v}_k\|}$. Now an eigenvalue is approximately $\frac{\|A\boldsymbol{v}_k\|}{\|\boldsymbol{v}_k\|}$ for some super large k. The larger the k, the better the approximation. ☺

So we iteratively apply A, and we normalize (i.e., set the vectors into unit vectors) at each step. Then we get a convergence behavior. This is called the power method, since we are essentially studying the power sequence of A. In practice, people sometimes just take $\boldsymbol{v} = \boldsymbol{e}_1$, so we are just looking at the first column of A^k for some large k.

With this idea in mind, here comes a multi-vector version of the power method.

Example 2.5.5. Suppose A is invertible.

Pick a random orthonormal basis $Q = \begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_n\end{bmatrix}$. Set $Q_0 = Q$. We shall let A acts on these vectors simultaneously, and then normalize simultaneously.

Whenever Q_k is defined, consider AQ_k, whose column form another basis. However, columns of AQ_k are no longer orthonormal! Therefore we perform a Gram-Schmidt orthogonalization, i.e., $AQ_k = Q_{k+1}R_{k+1}$. Here Q_{k+1} is the "simultaneously normalized" version of AQ_k, so we use this for the next iteration.

Hopefully Q_{k+1} converge towards something. It turns out that under nice conditions, it converge towards a matrix Q such that $A = QTQ^{\mathrm{T}}$ is the upper triangularization of A, and we can just read the diagonal of $Q^{\mathrm{T}}AQ$ to find ALL the eigenvalues simultaneously.

Finally, note that $R_{k+1} = Q_{k+1}^{\mathrm{T}}AQ_k \to Q^{\mathrm{T}}AQ = T$. So ASSUMING that R_k converges, then they must converge towards T. We can simply read the diagonal of R_k for super large k and it will be an approximation of diagonals of T.

(Remark: I have hidden away MANY assumptions of nice-ness. We do not delve into them because they are more suited for a numerical analysis class.) ☺

Finally we have reached the famous QR-algorithm, which is essentially the same as above.

Example 2.5.6. We do QR-decomposition (i.e., Gram-Schmidt orthogonalization) for A and get $A = Q_0R_0$. Then we do QR-decomposition for R_0Q_0 and get $R_0Q_0 = Q_1R_1$, and then we get $R_1Q_1 = Q_2R_2$ and so on. Then under nice-ness assumptions, R_k should converge towards the upper triangular T in the Schur decomposition $A = QTQ^{\mathrm{T}}$, and the diagonal entries converge towards all the eigenvalues. ☺

Why is this? This is essentially the previous example, where we started with the standard basis as our orthonormal basis. So $A(I) = Q_0R_0$. Now $AQ_0 = Q_0R_0Q_0 = Q_0Q_1R_1$. Then $A(Q_0Q_1) = Q_0R_0Q_0Q_1 = Q_0Q_1R_1Q_1 = Q_0Q_1Q_2R_2$ and it goes on like that. As you can see, the idea is the same. We iteratively apply A, and we simultanesouly normalize all columns at each step.

In general, we have $A(Q_0 \ldots Q_k) = (Q_0 \ldots Q_{k+1})R_{k+1}$, and hence (if possible) R_k converges towards the T in the Schur decomposition.

Of course there are situations where R_k fail to converge at all. Take a numerical analysis class if you want more details of this. This algorithm is nowadays a key step in finding eigenvalues and finding roots of polynomials.

2.6 Sylvestor's equation

There are many proofs of Jordan canonical form. Our proof here is essentially a geometric proof. We break down into invariant subspaces and yada yada done. There is also a very interesting (but less illuminating) algebraic proof, where we study polynomials and yada yada done. (Maybe I'll type up another optional section about this.)

Finally, here is a computational proof, using Schur decompositions, and row and column operations, we shall achive a block-diagonalization without using generalized eigenstuff.

First, by Schur decomposition, we can always upper triangularize a matrix. Here is a particularly interesting example

Example 2.6.1. Consider $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$. We know that it has eigenvalue 1 with algebraic multiplicity 2 and eigenvalue 2 with algebraic multiplicity 1.

Let us first try to put it in upper triangular form. When we do this, by picking the right filtration, we want to make sure that we are grouping eigenvalues of the same value together. So say we require the resulting upper triangular matrix to have diagonal 1,1,2.

Then first we need a vector $\boldsymbol{v}_1$ for eigenvalue 1, say $\boldsymbol{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. Note that we can do a (non-invariant) decomposition of the domain into $\mathbb{R}^3 = V_y \oplus V_{xz}$ where V_y represents the y-axis, while V_{xy} is the xz-plane. Then since $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix}$, the corresponding submaps of A would be $A_{y\to y} = \begin{bmatrix} 1 \end{bmatrix}$, $A_{y\to xz} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $A_{xz\to y} = \begin{bmatrix} -1 & 2 \end{bmatrix}$, $A_{xz\to xz} = \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}$.

So to continue our filtration, since we already have V_y chosen, we need to look at V_{xz} and thus the linear map $A_{xz\to xz}$. Let us find an eigenvector $\boldsymbol{v}_2$ of $A_{xz\to xz}$ for eigenvalue 1, say $\boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in V_{xz}$ (the coordinates here are under the basis $\boldsymbol{e}_1, \boldsymbol{e}_3$ for V_{xz}). Then it is in fact the unit vector in the z-axis, i.e., $\boldsymbol{v}_2 = \boldsymbol{e}_3$. You can check that $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is indeed A-invariant.

Now we already have $\boldsymbol{v}_1, \boldsymbol{v}_2$ chosen. To finish the filtration, we just need to pick any $\boldsymbol{v}_3$ that make this into a basis. Since we have $\boldsymbol{v}_1 = \boldsymbol{e}_2, \boldsymbol{v}_2 = \boldsymbol{e}_3$, we might as well just pick $\boldsymbol{v}_3 = \boldsymbol{e}_1$, and we are done.

Under the basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, we have A similar to $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 2 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] = \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix}$ with $A_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} 2 \end{bmatrix}$ and $B = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$. ☺

So by choosing the right filtration, our matrix is something like, say, $\begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix}$ where A_1 and A_2 has NO common eigenvalues. Now we would like to kill B to make this block diagonal. How to do this?

We want to perform $C \begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix} C^{-1}$ so that the resulting matrix is block diagonal. Note that since C here is invertible, it must corresponds to some row/column operations that must happen in pairs. Can we find the right row/column operation to do this?

Suppose $C = \begin{bmatrix} I & X \\ & I \end{bmatrix}$, i.e., it is a block operation. Then $C \begin{bmatrix} A_1 & B \\ & A_2 \end{bmatrix} C^{-1} = C \begin{bmatrix} A_1 & B + XA_2 - A_1X \\ & A_2 \end{bmatrix} C^{-1}$. So we need to find X such that $A_1X - XA_2 = B$ for given A_1, A_2, B. This is the Sylvester's equation.

Theorem 2.6.2. *Suppose A, B are $m \times m$ matrix and $n \times n$ matrix with no common eigenvalue. Then for any $m \times n$ matrix C, there is a UNIQUE solution X to the matrix equation $AX - XB = C$.*

Proof. First of all, let V be the space of all $m \times n$ matrices. Consider the map $L : V \to V$ such that $L(X) = AX - XB$. Note that, indeed, L would send an $m \times n$ matrix to another $m \times n$ matrix, and it is also linear! This means that it is an linear operator. Our goal is to show that L is a bijection, hence it is enough to check that the kernel of L is trivial.

So we have reduced our problem to this: we need to show that $AX - XB = 0$ must only have the solution $X = 0$. (See how the problem is simplified? THAT is why we do abstract vector spaces. We do not even need V, L from now on, but the abstraction allows us to SEE that we have a simplification.)

Suppose $AX - XB = 0$, then $AX = XB$. In particular, $A^k X = XB^k$ for any positive integer k. Now we take linear combinations of powers, we see that $p(A)X = Xp(B)$ for any polynomial $p(x)$.

Consider $p_A(x)$, the characteristic polynomial of A. Then on one hand, $p_A(A) = 0$. On the other hand, since A, B has no common eigenvalue, for each eigenvalue λ of B, $p_A(\lambda) \neq 0$. So $p_A(B)$ has NO eigenvalue zero. In particular, it is invertible! Hence we have $0 = p_A(A)X = Xp_A(B)$ where $p_A(B)$ is invertible, so $X = 0$ is the only solution. □

Example 2.6.3. We have A similar to $\left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] = \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix}$.

Now, since A_1 and A_2 has NO eigenvalue in common, we know that there is a unique $X \in M_{2\times 1}$ such that $A_1X - XA_2 = B$. Then A is similar to $\begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$.

To be more explicit, if $X = \begin{bmatrix} x \\ y \end{bmatrix}$, then $XA_2 - A_1X = -B$ would translate into $\begin{bmatrix} 2x \\ 2y \end{bmatrix} - \begin{bmatrix} x+2y \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$, which means that $x = -5$ and $y = -3$. Then A is similar to $\left[\begin{array}{cc|c} 1 & 0 & -5 \\ 0 & 1 & -3 \\ \hline 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cc|c} 1 & 2 & -1 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 2 \end{array}\right] \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & 3 \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$. This corresponds to a spatial decomposition of V into invariant subspaces.

Finally, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is a Jordan block, and $A_2 = (2)$ is already a Jordan block. So A is similar to $\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 2 & 0 \\ \hline 0 & 0 & 1 \end{array}\right] \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right] \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right]$ is block diagonal with Jordan blocks on the diagonal.

If you have been keeping track, we have done the Jordan canonical form of A with exclusively row/column operations that come in inverse paris, i.e., each step is $A \to XAX^{-1}$ for some elementary matrix X. ☺

2.7 (Optional) Polynomial proof of the Jordan canonical form

We quote without proof the following theorem. This is essentially the Euclidean algorithm for coprime polynomials. (Similar to the Euclidean algorithm for coprime numbers.)

Theorem 2.7.1. *If $p(x), q(x)$ have no common root, then we can find polynomials $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$.*

Proposition 2.7.2. *If $p(x), q(x)$ have no common root, then* $\mathrm{Ker}(p(A)q(A)) = \mathrm{Ker}(p(A)) \oplus \mathrm{Ker}(q(A))$.

Proof. Obviously $\mathrm{Ker}(p(A)), \mathrm{Ker}(q(A))$ are subspaces of $\mathrm{Ker}(p(A)q(A))$, so we just need to show that they are linearly independent and spanning.

If $\boldsymbol{v} \in \mathrm{Ker}(p(A)q(A))$, then $p(A)q(A)\boldsymbol{v} = \mathbf{0}$. In particular, $p(A)\boldsymbol{v} \in \mathrm{Ker}(q(A))$ and $q(A)\boldsymbol{v} \in \mathrm{Ker}(p(A))$.

Now since $p(x), q(x)$ have no common root, therefore we can find polynomials $a(x), b(x)$ such that $a(x)p(x) + b(x)q(x) = 1$. Then $a(A)p(A) + b(A)q(A) = I$. So $\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} \in \mathrm{Ker}(p(A)) + \mathrm{Ker}(q(A))$. So we have spanning.

Now let us show that the two subspaces have zero intersection. If $\boldsymbol{v} \in \mathrm{Ker}(p(A)) \cap \mathrm{Ker}(q(A))$, then $\boldsymbol{v} = a(A)p(A)\boldsymbol{v} + b(A)q(A)\boldsymbol{v} = \mathbf{0} + \mathbf{0} = \mathbf{0}$. So we are done. □

Corollary 2.7.3. *The whole domain is the direct sum of generalized eigenspaces.*

Proof. Note that $\mathrm{Ker}(p_A(A))$ is the whole domain for the characteristic polynomial $p_A(x)$ of A.

Decompose $p_A(x)$ as the product of coprime factors $p_i(x) = (x - \lambda_i)_i^m$. Then use the proposition above repeatedly. Done. □

Chapter 3

Functions of Matrices

3.1 Limit of Matrices

Whenever we have a collection of things, and a concept of "distance" between things, then we can define limits in the sense of shrinking distance. In this way, we can easily define limits of vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$. And treating matrices as vectors in $\mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$, we can define limits of matrices.

So, as an operational definition, one may think of the limit of a sequence of matrices $\{M_k\}_{k\in\mathbb{N}}$ as taking a limit on each entry.

Of course, technically speaking, this definition is bad. The "entries" of a matrix depend on your choice of basis. If you change basis, then all entries are now different. Who can guarantee that the limit will stay the same?

We can ad-hoc verify that this is the case.

Proposition 3.1.1. *If $\lim A_n = A$ and $\lim B_n = B$, then $\lim(A_nB_n)$ exists and it is AB.*

Proof. One line calculation proof. $\lim(\sum_k a_{ik,n}b_{kj,n}) = \sum_k \lim(a_{ik,n})\lim(b_{kj,n})$. □

Corollary 3.1.2. *$\lim(BA_nB^{-1}) = B(\lim A_n)B^{-1}$. So limits are invariant under a change of basis.*

But ad-hoc arguments are like cheating. A GOOD definition should make this clear in the first place. We do not require this good definition, but if you are curious, read the following remark.

Remark 3.1.3. *This is a exposition on how to define limits of linear operators without picking a basis. This portion is optional.*

A sequence of vectors in an abstract vector space has no well-defined limit. This is because there is no way to measure distance (or induce some topology), and therefore there is no way to measure convergence.

But with inner product structures, we are now golden. Given a sequence of vectors $\{\boldsymbol{v}_n\}_{n\in\mathbb{N}}$ in an inner product space V, we say their limit is $\boldsymbol{v}$ if for all $\epsilon > 0$, we can find $N \in \mathbb{N}$ such that $\|\boldsymbol{v} - \boldsymbol{v}_n\| < \epsilon$ whenever $n \geq N$. You know, the obvious way to define this.

Given linear maps $L, L' : V \to W$ between two inner product spaces, how to define distance? It turns out that there are many ways to define this distance. One would be the operator norm, where we define the norm $\|L\|$ to be the largest $\|L\boldsymbol{u}\|$ for all unit vectors $\boldsymbol{u}$. In particular, it is the largest possible length-dilation that can happen, $\max_{\boldsymbol{v}\in V}\frac{\|L\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$. This looks nice, yes? For any input $\boldsymbol{v}$, we shall always have $\|L\boldsymbol{v}\| \leq \|L\|\|\boldsymbol{v}\|$, and the norm $\|L\|$ is exactly the tightest possible constant k for $\|L\boldsymbol{v}\| \leq k\|\boldsymbol{v}\|$ to work for all $\boldsymbol{v}$.

It is even easy to conceptualize: it is exactly the largest singular value of L. (NOT the eigenvalue!) Neat!

Then using $\|L - L'\|$ as a distance between two linear maps, we can then define $L = \lim L_n$ in the obvious way. I.e., for all $\epsilon > 0$, we can find $N \in \mathbb{N}$ such that $\|L - L_n\| < \epsilon$ whenever $n \geq N$.

Note that our operator norm satisfy the condition that $\|LL'\| \leq \|L\|\|L'\|$. (Easy to prove as $\|LL'\boldsymbol{v}\| \leq \|L\|\|L'\boldsymbol{v}\| \leq \|L\|\|L'\|\|\boldsymbol{v}\|$.) As a result, if L_n converge to L and say L'_n converge to L', then $L_nL'_n$ would *converge to LL'. I.e., we have the identity $\lim(L_nL'_n) = (\lim L_n)(\lim L'_n)$ whenever the latter two limits exist.*

Now since matrix multiplication respect limits, if we pick orthonormal basis and assume that our domain and codomain are $\mathbb{C}^n, \mathbb{C}^m$, then we see that $\lim(\boldsymbol{e}_i^* L_n \boldsymbol{e}_j) = \boldsymbol{e}_i^*(\lim L_n)\boldsymbol{e}_j$. So if we picked some basis, then linear operator convergence is the same as convergence in all entries.

Let us define this norm in a different way. You may recall (or you can verify) that $\operatorname{trace}(L^*L')$ is an inner product of the space of linear maps from V to W, and we may define $\|L\|^2 = \operatorname{trace}(L^*L)$. To be more clear that this is independent of basis, we actually have $\|L\| = \sqrt{\sum \sigma_i^2}$ where σ_i are all the singular values. If we had picked an orthonormal basis, then we also have $\|L\| = \sqrt{\sum a_{ij}^2}$ where a_{ij} are all the entries. Neat right? This is a very natural way to define a norm, and it is NOT the same as the operator norm.

But worry not. You may verify that we still have $\|LL'\| \leq \|L\|\|L'\|$, and therefore we also have $\lim(L_nL'_n) = (\lim L_n)(\lim L'_n)$ and $\lim(\boldsymbol{e}_i^* L_n \boldsymbol{e}_j) = \boldsymbol{e}_i^*(\lim L_n)\boldsymbol{e}_j$.

So in the end, it does not matter much which norm we pick. The only important property here is $\|LL'\| \leq \|L\|\|L'\|$. As long as this condition is true, then the convergences in different settings mean exactly the same thing. The TOPOLOGY is the same.

Finally, above statements applies strictly to finite dimensional cases. For infinite dimensional spaces, the two norms above would induce different topologies and will have different meaning of convergence.

Now we have a TOPOLOGY (a way to talk about convergence) on matrices. Then we can define dense subsets.

Theorem 3.1.4. *Diagonalizable matrices are dense in $n \times n$ matrices. (I.e., any matrix is a limit of diagonalizable matrices.)*

对角矩阵是稠密的,和同实数是稠密一样,任意一个狭小区间邻域内都有对角阵

Proof. Given a matrix A, how to construct a sequence of diagonalizable matrices whose limit is A? First, we change basis and assume that A is in Jordan canonical form (or any upper triangular form).

Say the diagonal entries (eigenvalues) are $a_1, \ldots, a_n$. Note that some of these are the same, while some are not. Let g be the smallest "gap" between distinct diagonal entries, i.e., either $a_i = a_j$, or $|a_i - a_j| \geq g$.

For a tiny real number $t < \frac{g}{2n}$, consider a diagonal matrix $D(t) = \begin{bmatrix} t & & \\ & \ddots & \\ & & nt \end{bmatrix}$, let $A_t = A + D_t$. Then $\lim_{t\to 0} A_t = A$. I only need to show that A_t are diagonalizable.

Note that eigenvalues of A_t are $a_1 + t, \ldots, a_n + nt$. For any $i \neq j$, if $a_i = a_j$, then $a_i + it \neq a_j + jt$. If $|a_i - a_j| \geq g$, then $|(a_i + it) - (a_j + jt)| \geq g - it - jt \geq g - 2nt > 0$ by construction of t, so $a_i + it \neq a_j + jt$. Eitherway, we see that eigenvalues of A_t are all distinct, so it must be diagonalizable. Done. □

Note that we in fact proved something stronger: matrices with distinct eigenvalues are dense. Feel free to prove something even stronger: INVERTIBLE matrices with distinct eigenvalues are dense. (Just throw in distance to zero when you define the "gap" size g.)

This fact is extremely useful. Consider this:

余子式 $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \cdot \frac{1}{ad-bc}$

Corollary 3.1.5. *Given a square matrix A, let A_{ij} be its (i,j)-cofactor, and let $\operatorname{Adj}(A)$ be the adjugate matrix of A. (So for invertible matrices, $A^{-1} = \frac{1}{\det(A)}\operatorname{Adj}(A)$. Note that for non-invertible matrices, $\operatorname{Adj}(A)$ is still defined.)*

Then for any square matrices A, B, we havbe $\operatorname{Adj}(AB) = \operatorname{Adj}(B)\operatorname{Adj}(A)$.

Proof. Note that invertible matrices are dense. And for invertible matrices, $\operatorname{Adj}(AB) = \det(AB)(AB)^{-1} = \det(A)B^{-1}\det(A)A^{-1} = \operatorname{Adj}(B)\operatorname{Adj}(A)$. Now take limit and we are done. $\det(AB) = \det(A) \cdot \det(B)$ □

Or for example, let us prove Cayley-Hamilton again. First, if A has distinct eigenvalues, then it is trivial to verify that $p_A(A) = 0$. (A is diagonalizable, so there is a basis made of eigenvectors of A. And $p_A(A)$ will kill all eigenvectors of A.) Then by taking limits of A_n with distinct eigenvalues, we have $p_A(A)$ for any matrix A.

即 $P(A) \cdot X_A = 0$
而 $\forall X_A$ 加起来构成基

有不同特征值的 A_n 可行且其稠密
故 Any A 都成立

$A = P^{-1} J P$
$A - \lambda I = P^{-1}(J - \lambda I)P$
$(A-\lambda I)^{d_1} = P^{-1}(J-\lambda I)^{d_1}P$

分块阵化为0.
再相乘必为0

λ 代数重数为 d.

$\begin{matrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \lambda & 1 \\ & & & \lambda \end{matrix}$

Remark 3.1.6. *The adjugate matrix is NOT useful at all. It is an attempt to relate the matrix A with its inverse. However, the Cayley-Hamilton theorem does a better job at this.*

If A is invertible, then $\det(A) \neq 0$, so $p_A(x)$ has a non-zero constant term. I.e., $p_A(x) = xq(x) + a$ for some $a \neq 0$. Then since $p_A(A) = 0$, we have $Aq(A) + aI = 0$, and thus $A^{-1} = \frac{1}{a}q(A)$. So A^{-1} is ALWAYS a polynomial of A! $P(A)=(A-\lambda_1)^{d_1}\cdots$由于无0，故有$\lambda_1^{d_1}\cdots\lambda_n^{d_n}$ $=\det$

A polynomial relation is huge. Whatever you can do using adjugates, you can use Cayley-Hamilton instead. For example, a classical argument for the adjugate matrix goes like this: if entries of A are rational, then entries of A^{-1} are also rational. To see this, note that each cofactor is a sum of products of entries of A, so it is rational. So we are done. We also see that if A has integer entries, then $\det(A)A^{-1}$ has integer entries. 有理数

But with Cayley-Hamilton, $A^{-1} = \frac{1}{\det(A)}q(A)$, and the coefficients of $q(x)$ are also sums of products of entries of A. So if A has rational entries, then A^{-1} has rational entries, and if A has integer entries, then $\det(A)A^{-1}$ has integer entries.

3.2 Functions of matrices

What is a function of a matrix? Here is an easy example:

Definition 3.2.1. *We define e^A to be the limit $\lim_{n\to\infty}(I + A + \frac{1}{2!}A^2 + \cdots + \frac{1}{n!}A^n)$. This is the limit of a sequence of matrix.*

This raises an immediate problem. Why would this series converge at all? (Spoiler: it will always converge.) If we were in an analysis class, then we shall then proceed to show convergence. It is not too bad, as entries of A^n grows polynomially while the denominator $n!$ grows faster than exponential.

But as a linear algebra class, let us jump out of this, and think about something bigger. If YOU were to define a function of a matrix, $f(A)$ for some function f, what would you like?

The following principles seem like must-haves:

1. We want it to NOT depend on our choice of basis. So $f(BAB^{-1}) = Bf(A)B^{-1}$. It is really a function of LINEAR TRANSFORMATIONS.
2. We want it to respect independent actions. So $f(\begin{bmatrix} A & \\ & B \end{bmatrix}) = \begin{bmatrix} f(A) & \\ & f(B) \end{bmatrix}$. In particular, for diagonal matrices, $f(D)$ is just applying f on each diagonal entry.
3. If $f : \mathbb{R} \to \mathbb{R}$ or $f : \mathbb{C} \to \mathbb{C}$ is continuous, then the induced function $f : M_{n\times n} \to M_{n\times n}$ should still be continuous. Here $M_{n\times n}$ refers to the space of all $n \times n$ real or complex matrices, depending on context. (We can use real functions when all of our eigenvalues are real.)

Combining these principles, one thing is super clear. If A is diagonalizable $A = BDB^{-1}$ where $D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$, then we want $f(A) = Bf(D)B^{-1} = B\begin{bmatrix} f(d_1) & & \\ & \ddots & \\ & & f(d_n) \end{bmatrix}B^{-1}$. So this resulting matrix $f(A)$ is already uniquely defined! It is also not hard to see that, if A changes continuously (i.e., B and each d_i changes continuously), then since f is continuous on $\mathbb{C}$, $f(d_i)$ also changes continuously, and hence $f(A) = Bf(D)B^{-1}$ changes continuously. So we have all the desired result.

BUT what if A is NOT diagonalizable? This is where density comes into play. According to our principles, $f(\lim A_n) = \lim f(A_n)$. So we just use a sequence of diagonalizable matrices to approximate A, and we can get $f(A)$.

Would the limit always exists? Let us see what would happen.

Example 3.2.2. Consider $J = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$. Let $J_t = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda + t \end{bmatrix}$, then clearly J_t is diagonalizable whenever $t \neq 0$, and $\lim_{t\to 0} J_t = J$.

So for a function f, we want $f(J) = f(\lim_{t\to 0} J_t) = \lim_{t\to 0} f(J_t)$. To calculate $f(J_t)$, we need to diagonalize J_t. Note that $J_t = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda + t \end{bmatrix}$ can be diagonalized by solving the corresponding sylvesters equation $\lambda x - x(\lambda + t) = 1$, which yeilds $x = -\frac{1}{t}$. So $\begin{bmatrix} 1 & \frac{1}{t} \\ 0 & 1 \end{bmatrix} J_t \begin{bmatrix} 1 & -\frac{1}{t} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda & \\ & \lambda + t \end{bmatrix}$. In particular, we have $J_t = \begin{bmatrix} 1 & \frac{1}{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda & \\ & \lambda + t \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{t} \\ 0 & 1 \end{bmatrix}$.

So $f(J) = f(\lim_{t\to 0} J_t) = \lim_{t\to 0} f(J_t) = \lim_{t\to 0} \begin{bmatrix} 1 & \frac{1}{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f(\lambda) & \\ & f(\lambda + t) \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{t} \\ 0 & 1 \end{bmatrix} = \lim_{t\to 0} \begin{bmatrix} f(\lambda) & \frac{f(\lambda+t)-f(\lambda)}{t} \\ & f(\lambda + t) \end{bmatrix}$.

Wait, the definition of the derivative is right there!

So we must have $f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) \\ & f(\lambda) \end{bmatrix}$ when f is differentiable at λ. Otherwise $f(J)$ cannot be defined and $\lim f(J_t)$ does not converge. ☺

This line of logic can easily be generalized to give us a formula for $f(A)$ in general. But for mnemonics sake, let us see an alternative proof.

Proposition 3.2.3. *Assume that f is analytical at λ. (It means f equals to its Taylor expansion at λ.) Consider the $n \times n$ Jordan block* $J = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$. *Then using previous principles, we must have*

$$f(J) = \begin{bmatrix} f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(n-1)!}f^{(n-1)}(\lambda) \\ & \ddots & \ddots & \vdots \\ & & \ddots & \frac{1}{1!}f'(\lambda) \\ & & & f(\lambda) \end{bmatrix}.$$

Proof. Why do we see coefficients of Taylor expansions? That is not a coincidence. First we have $J = N + \lambda I$ where N is the nilpotent Jordan block.

Now f equals to its Taylor series. So if we expand f at λ, we have $f(x) = a_0 + a_1(x-\lambda) + a_2(x-\lambda)^2 + \dots$ where $a_k = \frac{1}{k!}f^{(k)}(\lambda)$. So $f(J) = a_0 I + a_1 N + a_2 N^2 + \dots$. But as a nilpotent matrix, $N^n = 0$, and N^k is really just the identity matrix shifted up k times. So $f(J) = a_0 I + a_1 N + a_2 N^2 + \dots + a_{n-1}N^{n-1} = \begin{bmatrix} f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(n-1)!}f^{(n-1)}(\lambda) \\ & \ddots & \ddots & \vdots \\ & & \ddots & \frac{1}{1!}f'(\lambda) \\ & & & f(\lambda) \end{bmatrix}$. □

Now we can define functions of matrices.

Definition 3.2.4. *Suppose f is a function defined at all eigenvalues of A, and it is $(m-1)$-times differentiable at the eigenvalue λ when λ-blocks in the Jordan canonical form of A have sizes at most m. Then we define $f(J)$ for each involved Jordan blck as*

$$f(J) = \begin{bmatrix} f(\lambda) & \frac{1}{1!}f'(\lambda) & \dots & \frac{1}{(m-1)!}f^{(m-1)}(\lambda) \\ & \ddots & \ddots & \vdots \\ & & \ddots & \frac{1}{1!}f'(\lambda) \\ & & & f(\lambda) \end{bmatrix}$$

, and we define

$$f(A) = B \begin{bmatrix} f(J_1) & & \\ & \ddots & \\ & & f(J_t) \end{bmatrix} B^{-1}$$

where the Jordan decomposition of A is $B \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_t \end{bmatrix} B^{-1}$.

Here is an obvious result:

足够光滑

Corollary 3.2.5. *If f is infinitely differentiable everywhere, then $f(A)$ is defined for all square matrix A.*

Corollary 3.2.6. *If A has eigenvalues $\lambda_1, \dots, \lambda_n$ counting algebraic multiplicity, then $f(A)$ has eigenvalues $f(\lambda_1), \dots, f(\lambda_n)$ counting algebraic multiplicity.*

Proof. Do it block-wise. ☐

Corollary 3.2.7. *$f(A)g(A) = h(A)$ if $f(x)g(x) = h(x)$, and $f(A) + g(A) = h(A)$ if $f(x) + g(x) = h(x)$. Finally, if $f(x) = x$, then $f(A) = A$, and if $f = 1$ is a constant function, then $f(A) = I$.*

Proof. Do it block-wise. $f(x) = x^3 + x^2 + x + 1$
$f(A) = A^3 + A^2 + A + I$ ☐

Corollary 3.2.8. *If f is a polynomial, then $f(A)$ is exactly as we have always defined it to be.*

One can then do the boring verification that such a definition satisfy the given principles. We are going to skip those because we might not learn much from that process.

Corollary 3.2.9. *If $f = g$ at all eigenvalues of A and they also equal at enough derivatives that are used in $f(A)$ and $g(A)$, then $f(A) = g(A)$.*

Corollary 3.2.10. *Fix a matrix A, then for any well-defined $f(A)$, there is a polynomial $p(x)$ such that $f(A) = p(A)$.* 定义多项式

Proof. Let us say the largest Jordan block in A has size m, and eigenvalues are $\lambda_1, \dots, \lambda_k$. Then we can always find a polynomial whose j-th derivatives at λ_i is some prescribed value, for all $i \leq k$ and all $j < m$. (In fact we can also require the degree of this polynomial to be at most $(m-1)j$.) ☐

So if we are FIXING A, then there is NO point in studying $f(A)$ at all. They are all just polynomials of A. In particular, we have results like $Af(A) = f(A)A$ always, and etc.

However, be careful here. If we are fixing f, but changing A, then each different A might require a different polynomial. So it is better to study $f(A)$ in terms of f.

Example 3.2.11. For $A = I$, then $\mathrm{e}^A = \mathrm{e}A$, so $f(A) = p(A)$ where $p(x) = \mathrm{e}x$.

But for $A = \begin{bmatrix} 0 & 1 \\ & 0 \end{bmatrix}$, then $\mathrm{e}^A = I + A = q(A)$ where $q(x) = x + 1$. ☺

3.3 Applications to functions of Matrices

The obvious application is to solve various differential equations.

Lemma 3.3.1. *If $AB = BA$, then $\mathrm{e}^{A+B} = \mathrm{e}^A \mathrm{e}^B = \mathrm{e}^B \mathrm{e}^A$.*

Proof. Direct computation using Taylor series of e^x. (But don't miss out on the alternative conceptual proof!)

$$e^A e^B = (\sum_m \frac{1}{m!}A^m)(\sum_n \frac{1}{n!}B^n) = \sum_{m,n} \frac{1}{m!n!}A^m B^n.$$

Now let $k = m+n$. We have

$$\sum_{m,n} \frac{1}{m!n!}A^m B^n = \sum_k \sum_{n=0}^{k} \frac{1}{n!(k-n)!}A^(k-n)B^n = \sum_k \frac{1}{k!} \sum_{n=0}^{k} \frac{k!}{n!(k-n)!}A^(k-n)B^n = \sum_k \frac{1}{k!}(A+B)^k.$$

Note that commutativity $AB = BA$ is used in the last step. For example, $A^2 + 2AB + B^2 = (A+B)^2$ is only true when we have commutativity. □

Proof. By density, it is enough to prove this when A has distinct eigenvalues. Then $AB = BA$ imples that A, B are simultaneously diagonalizable. So we may assume that A, B are both diagonal, and then the statement is trivial. □

Remark 3.3.2. *When A has distinct eigenvalues, then $AB = BA$ imples that A, B are simultaneously diagonalizable. Hopefully your last linear algebra class has discussed this. But if not, see if you can prove this yourself.*

There are many proofs. If you need a hint, maybe try 2 by 2 matrices. If $\begin{bmatrix} a & \\ & b \end{bmatrix} X = X \begin{bmatrix} a & \\ & b \end{bmatrix}$ *and $a \neq b$, whwy must X be diagonal?*

Or one can work abstractly on finding common eigenvectors.

Proposition 3.3.3. $\frac{d}{dt}e^{At} = Ae^{At}$.

Proof. Compute. Or be cheap and do this for diagonal A, and use density. □

Corollary 3.3.4. *Let $\boldsymbol{v}(t)$ be a vector of functions, i.e., each coordinate may change as t change. Suppose it satisfy the differential equation $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$ for some linear transformation A. Then $e^{At}\boldsymbol{c}$ is a solution for any constant vector $\boldsymbol{c}$. (In fact $\boldsymbol{v}(0) = \boldsymbol{c}$, so it is the initial condition.)*

I claim that this is in fact the only solution.

Proposition 3.3.5. *The solution space to $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$ is n dimensional where n is the dimension of the domain. (So columns of e^{At} form a basis.)*

Proof. Let us first do a single nilpotent Jordan block. Then we have $\begin{bmatrix} f_1' \\ \vdots \\ \vdots \\ f_n' \end{bmatrix} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ \vdots \\ f_n \end{bmatrix}$. This reads as $f_i' = f_{i+1}$. So the solution space is simply this: f_1 is a polynomial of degree at most $n-1$, and the rest are iterated derivatives of f_1. Obviously the solution space is n dimensional.

Now let us do a single λ-Jordan block J. Let $\boldsymbol{w}(t) = e^{-\lambda t}\boldsymbol{v}(t)$. Then $\boldsymbol{w}'(t) = e^{-\lambda t}\boldsymbol{v}'(t) - \lambda e^{-\lambda t}\boldsymbol{v}(t) = e^{-\lambda t}(J - \lambda I)\boldsymbol{v}(t) = (J - \lambda I)\boldsymbol{w}(t)$. But $(J - \lambda I)$ is nilpotent, so solutions to $\boldsymbol{w}(t)$ is n dimensional. Hence $\boldsymbol{v}(t) = e^{\lambda t}\boldsymbol{w}(t)$ has n-dimensions of possibilities as well.

Now suppose A has many Jordan blocks. But being block diagonal means each block behaves independently, so we are reduced to the single block cases and we are done. □

Conclusion: given a differential equation $\boldsymbol{v}'(t) = A\boldsymbol{v}(t)$ and initial value $\boldsymbol{v}(0) = \boldsymbol{c}$, then the unique solution is $e^{At}\boldsymbol{c}$.

You can imagine that things like $\sin(A)$ and such will also help solving other kinds of differential equations. We leave the rest to your future differential equation class.

Let us see another use of functions of matrices.

Definition 3.3.6. *We define the sign function* sign *such that* $\mathrm{sign}(a+b\mathrm{i}) = 1$ *if* $a > 0$*,* $\mathrm{sign}(a+b\mathrm{i}) = -1$ *if* $a < 0$*, and undefined when* $a = 0$*.*

It is obvious that this sign function is smooth (infinitely differentiable) whenever the input is NOT purely imaginary. So for any matrix A whose eigenvalues are NOT purely imaginary, then $\mathrm{sign}(A)$ is well-defined. Specifically, for any λ-Jordan block J, then $\mathrm{sign}(J) = \mathrm{sign}(\lambda)I = \pm I$.

Let us consider an application of this sign function.

Example 3.3.7. Consider the following variants of the Sylvester's equation. We want to find X to solve $AX + XB = C$, where A, B have positive eigenvalues.

(This is very possible, because in physics, eigenvalues are usually energy states or some other physical meanings, which we usually want to be positive.)

Solving this equation is the same as finding a diagonalization $\begin{bmatrix} A & -C \\ & -B \end{bmatrix} = \begin{bmatrix} I & X \\ & I \end{bmatrix} \begin{bmatrix} A & \\ & -B \end{bmatrix} \begin{bmatrix} I & -X \\ & I \end{bmatrix}$.

Now apply matrix sign function and watch the magic:

$\mathrm{sign}(\begin{bmatrix} A & -C \\ & -B \end{bmatrix}) = \begin{bmatrix} I & X \\ & I \end{bmatrix} \mathrm{sign}(\begin{bmatrix} A & \\ & -B \end{bmatrix}) \begin{bmatrix} I & -X \\ & I \end{bmatrix} = \begin{bmatrix} I & X \\ & I \end{bmatrix} \begin{bmatrix} I & \\ & -I \end{bmatrix} \begin{bmatrix} I & -X \\ & I \end{bmatrix} = \begin{bmatrix} I & -2X \\ & -I \end{bmatrix}$.

So if we have a magic computer to compute matrix sign function, then to solve $AX + XB = C$, we simply apply the matrix sign function to $\begin{bmatrix} A & -C \\ & -B \end{bmatrix}$ and read the answer from the upper right block. ☺

Remark 3.3.8. *(This part should be moved to earlier sections....)*

The Sylvester's equations are very important. For example, consider the case $AX - XB = C$ *where* $C = 0$ *and* B *is* 1×1*. Then we have* $AX = Xb$ *for some number* b*, and* X *is* $m \times 1$*, a vector! In particular, this is the equation defining eigenvectors and eigenvalues. In general, for the equation* $AX = XB$*, you may think of the solution* X *as the* B*-eigenstuff for* A*. And* $AX - XB = C$ *is the inhomogeneous version of this. (Just like how* $f' - f = 0$ *and* $f' - f = x^2$ *are related.)*

Furthermore, if $AX = XB$*, then* $\mathrm{Ran}(X)$ *is an invariant subspace of* A*. Can you see this? (And* $\mathrm{Ran}(X^{\mathrm{T}})$ *is an invariant space of* B^{T}*.)*

3.4 Matrix exponentials, curves and rotations