

# 理解矩阵（一） - 孟岩 - 博客频道

版权声明：本文为博主原创文章，未经博主允许不得转载。

前不久chensh出于不可告人的目的，要充当老师，教别人线性代数。于是我被揪住就线性代数中一些务虚性的问题与他讨论了几次。很明显，chensh觉得，要让自己在讲线性代数的时候不被那位强势的学生认为是神经病，还是比较难的事情。可怜的chensh，谁让你趟这个地雷阵？！色令智昏啊！线性代数课程，无论你是否从行列式入手还是直接从矩阵入手，从一开始就充斥着莫名其妙。比如说，在全国一般工科院系教学中应用最广泛的同济线性代数教材（现在到了第四版），一上来就介绍逆序数这个“前无古人，后无来者”的古怪概念，然后用逆序数给出行列式的一个极不直观的定义，接着是一些简直犯傻的行列式性质和习题——把这行乘一个系数加到另一行上，再把那一列减过来，折腾得那叫一个热闹，可就是压根看不出这个东西有嘛用。大多数像我一样资质平庸的学生到这里就有点犯晕：连这是个什么东西都模模糊糊的，就开始钻火圈表演了，这未免太“无厘头”了吧！于是开始有人逃课，更多的人开始抄作业。这下就中招了，因为其后的发展可以用一句峰回路转来形容，紧跟着这个无厘头的行列式的，是一个同样无厘头但是伟大的无以复加的家伙的出场——矩阵来了！多年之后，我才明白，当老师犯傻似地用中括号把一堆傻了吧叽的数据括起来，并且不紧不慢地说：“这个东西叫做矩阵”的时候，我的数学生涯掀开了何等悲壮辛酸、惨绝人寰的一幕！自那以后，在几乎所有跟“学问”二字稍微沾点边的东西里，矩阵这个家伙从不缺席。对于我这个没能一次搞定线性代数的笨蛋来说，矩阵老大的不请自来每每搞得我灰头土脸，头破血流。长期以来，我在阅读中一见矩阵，就如同阿Q见到了假洋鬼子，揉揉额角就绕道走。事实上，我并不是特例。一般工科学生初学线性代数，通常都会感到困难。这种情形在国内外皆然。瑞典数学家Lars Garding在其名著Encounter with Mathematics中说：“如果不熟悉线性代数的概念，要去学习自然科学，现在看来就和文盲差不多。”，然而“按照现行的国际标准，线性代数是公理化来表述的，它是第二代数学模型，...，这就带来了教学上的困难。”事实上，当我们开始学习线性代数的时候，不知不觉就进入了“第二代数学模型”的范畴当中，这意味着数学的表述方式和抽象性有了一次全面的进化，对于从小一直在“第一代数学模型”，即以实用为导向的、具体的数学模型中学习的我们来说，在没有并明确告知的情况下进行如此剧烈的paradigm shift，不感到困难才是奇怪的。大部分工科学生，往往是在学习了一些后继课程，如数值分析、数学规划、矩阵论之后，才逐渐能够理解和熟练运用线性代数。即便如此，不少人即使能够很熟练地以线性代数为工具进行科研和应用工作，但对于很多这门课程的初学者提出的、看上去是很基础的问题却并不清楚。比如说：\* 矩阵究竟是什么东西？向量可以被认为是具有n个相互独立的性质（维度）的对象的表示，矩阵又是什么呢？我们如果认为矩阵是一组列（行）向量组成的新的复合向量的展开式，那么为什么这种展开式具有如此广泛的应用？特别是，为什么偏偏二维的展开式如此有用？如果矩阵中每一个元素又是一个向量，那么我们再展开一次，变成三维的立方阵，是不是更有用？\* 矩阵的乘法规则究竟为什么这样规定？为什么这样一种怪异的乘法规则却能够在实践中发挥如此巨大的功效？很多看上去似乎是完全不相关的问题，最后竟然都归结到矩阵的乘法，这难道不是很奇妙的事情？难道在矩阵乘法那看上去莫名其妙的规则下面，包含着世界的某些本质规律？如果是的话，这些本质规律是什么？\* 行列式究竟是一个什么东西？为什么会有如此怪异的计算规则？行列式

与其对应方阵本质上是有什么关系？为什么只有方阵才有对应的行列式，而一般矩阵就没有（不要觉得这个问题很蠢，如果有必要，针对 $m \times n$ 矩阵定义行列式不是做不到的，之所以不做，是因为没有这个必要，但是为什么没有这个必要）？而且，行列式的计算规则，看上去跟矩阵的任何计算规则都没有直观的联系，为什么又在很多方面决定了矩阵的性质？难道这一切仅是巧合？\* 矩阵为什么可以分块计算？分块计算这件事情看上去是那么随意，为什么竟是可行的？\* 对于矩阵转置运算 $A^T$ ，有 $(AB)^T = B^T A^T$ ，对于矩阵求逆运算 $A^{-1}$ ，有 $(AB)^{-1} = B^{-1} A^{-1}$ 。两个看上去完全没有什么关系的运算，为什么有着类似的性质？这仅仅是巧合吗？\* 为什么说 $P^{-1}AP$ 得到的矩阵与 $A$ 矩阵“相似”？这里的“相似”是什么意思？\* 特征值和特征向量的本质是什么？它们定义就让人很惊讶，因为 $Ax = \lambda x$ ，一个诺大的矩阵的效应，竟然不过相当于一个小小的数 $\lambda$ ，确实有点奇妙。但何至于用“特征”甚至“本征”来界定？它们刻划的究竟是什么？这样的一类问题，经常让使用线性代数已经很多年的人都感到为难。就好像大人面对小孩子的刨根问底，最后总会迫不得已地说“就这样吧，到此为止”一样，面对这样的问题，很多老手们最后也只能用：“就是这么规定的，你接受并且记住就好”来搪塞。然而，这样的问题如果不能获得回答，线性代数对于我们来说就是一个粗暴的、不讲道理的、莫名其妙的规则集合，我们会感到，自己并不是在学习一门学问，而是被不由分说地“抛到”一个强制的世界中，只是在考试的皮鞭挥舞之下被迫赶路，全然无法领略其中的美妙、和谐与统一。直到多年以后，我们已经发觉这门学问如此的有用，却仍然会非常迷惑：怎么这么凑巧？我认为，这是我们的线性代数教学中直觉性丧失的后果。上述这些涉及到“如何能”、“怎么会”的问题，仅仅通过纯粹的数学证明来回答，是不能令提问者满意的。比如，如果你通过一般的证明方法论证了矩阵分块运算确实可行，那么这并不能够让提问者的疑惑得到解决。他们真正的困惑是：矩阵分块运算为什么竟然是可行的？究竟只是凑巧，还是说这是由矩阵这种对象的某种本质所必然决定的？如果是后者，那么矩阵的这些本质是什么？只要对上述那些问题稍加考虑，我们就会发现，所有这些问题都不是单纯依靠数学证明所能够解决的。像我们的教科书那样，凡事用数学证明，最后培养出来的学生，只能熟练地使用工具，却欠缺真正意义上的理解。自从1930年代法国布尔巴基学派兴起以来，数学的公理化、系统性描述已经获得巨大的成功，这使得我们接受的数学教育在严谨性上大大提高。然而数学公理化的一个备受争议的副作用，就是一般数学教育中直觉性的丧失。数学家们似乎认为直觉性与抽象性是矛盾的，因此毫不犹豫地牺牲掉前者。然而包括我本人在内的很多人都对此表示怀疑，我们不认为直觉性与抽象性一定相互矛盾，特别是在数学教育中和数学教材中，帮助学生建立直觉，有助于它们理解那些抽象的概念，进而理解数学的本质。反之，如果一味注重形式上的严格性，学生就好像被迫进行钻火圈表演的小白鼠一样，变成枯燥的规则之奴隶。对于线性代数的类似上述所提到的一些直觉性的问题，两年多来我断断续续地反复思考了四、五次，为此阅读了好几本国内外线性代数、数值分析、代数和数学通论性书籍，其中像前苏联的名著《数学：它的内容、方法和意义》、龚昇教授的《线性代数五讲》、前面提到的Encounter with Mathematics（《数学概观》）以及Thomas A. Garrity的《数学拾遗》都给我很大的启发。不过即使如此，我对这个主题的认识也经历了好几次自我否定。比如以前思考的一些结论曾经写在自己的blog里，但是现在看来，这些结论基本上都是错误的。因此打算把自己现在的有关理解比较完整地记录下来，一方面是因为我觉得现在的理解比较成熟了，可以拿出来与别人探讨，向别人请教。另一方面，如果以后再有进一步的认识，把现在的理解给推翻了，那现在写的这个snapshot也是很有意义的。因为打算写得比较多，所以会分几次慢慢写。也不知道是不是有时间

慢慢写完整，会不会中断，写着看吧。

---

今天先谈谈对线形空间和矩阵的几个核心概念的理解。这些东西大部分是凭着自己的理解写出来的，基本上不抄书，可能有错误的地方，希望能够被指出。但我希望做到直觉，也就是说能把数学背后说的实质问题说出来。首先说说空间(space)，这个概念是现代数学的命根子之一，从拓扑空间开始，一步步往上加定义，可以形成很多空间。线形空间其实还是比较初级的，如果在里面定义了范数，就成了赋范线性空间。赋范线性空间满足完备性，就成了巴那赫空间；赋范线性空间中定义角度，就有了内积空间，内积空间再满足完备性，就得到希尔伯特空间。总之，空间有很多种。你要是去看某种空间的数学定义，大致都是“存在一个集合，在这个集合上定义某某概念，然后满足某些性质”，就可以被称为空间。这未免有点奇怪，为什么要用“空间”来称呼一些这样的集合呢？大家将会看到，其实这是很有道理的。我们一般人最熟悉的空间，毫无疑问就是我们生活在其中的（按照牛顿的绝对时空观）的三维空间，从数学上说，这是一个三维的欧几里德空间，我们先不管那么多，先看看我们熟悉的这样一个空间有些什么最基本的特点。仔细想想我们就会知道，这个三维的空间：1. 由很多（实际上是无穷多个）位置点组成；2. 这些点之间存在相对的关系；3. 可以在空间中定义长度、角度；4. 这个空间可以容纳运动，这里我们所说的运动是从一个点到另一个点的移动（变换），而不是微积分意义上的“连续”性的运动，上面的这些性质中，最最关键的是第4条。第1、2条只能说是空间的基础，不算是空间特有的性质，凡是讨论数学问题，都得有一个集合，大多数还得在这个集合上定义一些结构（关系），并不是说有了这些就算是空间。而第3条太特殊，其他的空间不需要具备，更不是关键的性质。只有第4条是空间的本质，也就是说，容纳运动是空间的本质特征。认识到了这些，我们就可以把我们关于三维空间的认知扩展到其他的空间。事实上，不管是什么空间，都必须容纳和支持在其中发生的符合规则的运动（变换）。你会发现，在某种空间中往往会存在一种相对应的变换，比如拓扑空间中有拓扑变换，线性空间中有线性变换，仿射空间中有仿射变换，其实这些变换都只不过是空间允许的对应运动形式而已。因此只要知道，“空间”是容纳运动的一个对象集合，而变换则规定了对应空间的运动。下面我们来看看线性空间。线性空间的定义任何一本书上都有，但是既然我们承认线性空间是个空间，那么有两个最基本的问题必须首先得到解决，那就是：1. 空间是一个对象集合，线性空间也是空间，所以也是一个对象集合。那么线性空间是什么样的对象的集合？或者说，线性空间中的对象有什么共同点吗？2. 线性空间中的运动如何表述的？也就是，线性变换是如何表示的？我们先来回答第一个问题，回答这个问题的时候其实是不用拐弯抹角的，可以直截了当的给出答案。线性空间中的任何一个对象，通过选取基和坐标的办法，都可以表达为向量的形式。通常的向量空间我就不说了，举两个不那么平凡例子：L1. 最高次项不大于n次的多项式的全体构成一个线性空间，也就是说，这个线性空间中的每一个对象是一个多项式。如果我们以 $x^0, x^1, \dots, x^n$ 为基，那么任何一个这样的多项式都可以表达为一组n+1维向量，其中的每一个分量 $a_i$ 其实就是多项式中 $x^{(i-1)}$ 项的系数。值得说明的是，基的选取有多种办法，只要所选取的那一组基线性无关就可以。这要用到后面提到的概念了，所以这里先不说，提一下而已。L2. 闭区间 $[a, b]$ 上的n阶连续可微函数的全体，构成一个线性空间。也就是说，这个线性空间的每一个对象是一个连续函数。对于其中任何一个连续函数，根据魏尔斯特拉斯定理，一定可以找到最高次项不大于n的多项式函数，使之与该连续函数的

差为0，也就是说，完全相等。这样就把问题归结为L1了。后面就不用再重复了。所以说，向量是很厉害的，只要你找到合适的基，用向量可以表示线性空间里任何一个对象。这里头大有文章，因为向量表面上只是一列数，但是其实由于它的有序性，所以除了这些数本身携带的信息之外，还可以在每个数的对应位置上携带信息。为什么在程序设计中数组最简单，却又威力无穷呢？根本原因就在于此。这是另一个问题了，这里就不说了。下面来回答第二个问题，这个问题的回答会涉及到线性代数的一个最根本的问题。线性空间中的运动，被称为线性变换。也就是说，你从线性空间中的一个点运动到任意的另外一个点，都可以通过一个线性变化来完成。那么，线性变换如何表示呢？很有意思，在线性空间中，当你选定一组基之后，不仅可以用一个向量来描述空间中的任何一个对象，而且可以用矩阵来描述该空间中的任何一个运动（变换）。而使某个对象发生对应运动的方法，就是用代表那个运动的矩阵，乘以代表那个对象的向量。简而言之，在线性空间中选定基之后，向量刻画对象，矩阵刻画对象的运动，用矩阵与向量的乘法施加运动。是的，矩阵的本质是运动的描述。如果以后有人问你矩阵是什么，那么你就可以响亮地告诉他，矩阵的本质是运动的描述。（chensh，说你呢！）可是多么有意思啊，向量本身不是也可以看成是 $n \times 1$ 矩阵吗？这实在是很奇妙，一个空间中的对象和运动竟然可以用相类同的方式表示。能说这是巧合吗？如果是巧合的话，那可真是幸运的巧合！可以说，线性代数中大多数奇妙的性质，均与这个巧合有直接的关系。（待续）