Diagonalization of integer matrices

Linear algebra

Definition 1

Any group isomorphic to $\mathbb{Z}^n$, for some $n \in \mathbb{N}$, is called a **free abelian group of rank** *n.*

The group $\mathbb{Z}^n$ is analogous to the vector space $\mathbb{R}^n$; for example, the definitions of rank coincide. In this spirit, we shall call n-tuples in $\mathbb{Z}^n$ **vectors**.

Definition 2

Let G denote a finitely generated abelian group and suppose $S = \{\mathbf{g}_1, \ldots, \mathbf{g}_k\} \subset G$.

(a) *S is* **independent** *means if*

$$a_1\mathbf{g}_1 + \cdots + a_k\mathbf{g}_k = 0 \text{ with } a_i \in \mathbb{Z} \text{ for all } i = 1, \ldots, k, \tag{1.1}$$

then $a_i = 0$ for all $i = 1, \ldots, k$.

(b) *Any instance where Equation (1.1) fails is called a* **relation***.*

(c) *S is a* **basis** *means it is independent and generates G.*

The vectors $\mathbf{e}_i$ defined in Section **??** are called **standard basis vectors**.

Exercise 1

Let $G \cong \mathbb{Z}^n$. Use Equation (1.1) to show that if $S \subset G$ forms a basis for G then every vector $\mathbf{g} \in G$ has a unique expression as a ($\mathbb{Z}$)-linear combination of the elements in S.

A finitely generated abelian group has a basis if and only if it is free.

Using bases, we have a concise way to express homomorphisms between (finitely generated) free abelian groups.

Suppose $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a basis for $G \cong \mathbb{Z}^n$ and $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ is a basis for $H \cong \mathbb{Z}^m$. Let $\varphi : G \to H$ denote a group homomorphism such that

$$\varphi(\mathbf{b}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{c}_i, \quad j = 1, \ldots, n$$

for integers a_{ij}. Let A denote the $m \times n$ integer matrix whose (i,j)th entry is a_{ij} for $i = 1, \ldots, m$ and $j = 1 \ldots, n$.

By Exercise 1, every element in G can be written $\mathbf{g} = g_1\mathbf{b}_1 + \cdots g_n\mathbf{b}_n$ for unique $g_1, \ldots, g_n \in \mathbb{Z}$.

Let us abuse notation and write $\mathbf{g} = (g_1, \ldots, g_n) = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}$.

In this way we can write

$$\varphi : \mathbf{g} \mapsto \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}g_1 + a_{12}g_2 + \cdots + a_{1n}g_n \\ a_{21}g_1 + a_{22}g_2 + \cdots + a_{2n}g_n \\ \vdots \\ a_{m1}g_1 + a_{m2}g_2 + \cdots + a_{mn}g_n \end{pmatrix} = A\mathbf{g}.$$

In other words, the homomorphism φ is left multiplication by A.

Summary: Upon choosing bases for $G \cong \mathbb{Z}^n$ and $H \cong \mathbb{Z}^m$, we can express $\varphi : G \to H$ using the matrix A. We simply define φ by assigning an image to each of the basis elements of G... the caveat, of course, is the dependence on choice of bases.

Question

How do we get around this caveat?

Normalizing

Recall: In Linear Algebra, given bases $\mathfrak{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\} \subset \mathbb{R}^n$ and $\mathfrak{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \subset \mathbb{R}^n$ we have a **change of basis matrix** obtained by writing the vectors $\mathbf{b}_j$ ($j = 1, \ldots, n$) as linear combinations of the vectors $\mathbf{c}_i$ ($i = 1, \ldots, n$). The matrix is $U = (u_{ij})$, where

$$\mathbf{b}_j = \sum_{i=1}^{n} u_{ij}\mathbf{c}_i, \quad j = 1, \ldots, n.$$

Note, U is necessarily invertible. Left multiplication by U expresses vectors with respect to $\mathfrak{B}$ as vectors with respect to $\mathfrak{C}$; given $\mathbf{v} = \sum_{j=1}^{n} v_j\mathbf{b}_j$ put $\mathbf{w} := U\mathbf{v}$. Then $U^{-1}\mathbf{w} = \mathbf{v}$.

Example 1

Suppose $\mathbb{R}^n$ has basis $\mathfrak{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ and $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation given, with respect to $\mathfrak{B}$, by the matrix B. For $\mathbf{v} = \sum_{j=1}^n v_j \mathbf{b}_j$, define

$$\varphi_{\mathfrak{B}} : \mathbf{v} \mapsto B\mathbf{v}.$$

Suppose $\mathfrak{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ is also a basis for $\mathbb{R}^n$ and we have a change of basis matrix from $\mathfrak{B}$ to $\mathfrak{C}$ given by $U = (u_{ij})$. Because a change of basis is an isomorphism from $\mathbb{R}^n$ to $\mathbb{R}^n$ (also called an **automorhpism** on $\mathbb{R}^n$), for $\mathbf{w} = \sum_{i=1}^n w_i \mathbf{c}_i \in \mathbb{R}^n$, we may write $\mathbf{v} = U^{-1}\mathbf{w}$ for some $\mathbf{v} \in \mathbb{R}^n$. Then

$$\varphi_{\mathfrak{C}} : \mathbf{w} \mapsto UB\mathbf{v} = UBU^{-1}\mathbf{w}$$

gives the same transformation φ, but with respect to $\mathfrak{C}$.

The conjugation of B by U is equivalent to performing row and column operations on B. Left multiplication by U performs the row operations and right multiplication by U^{-1} performs the column operations. Furthermore, there is always a basis $\mathfrak{E}$ such that $\varphi_{\mathfrak{E}}$ is given by (left) multiplication by a **normalized** matrix E, a matrix consisting of zeros everywhere except for 1s on the first $r \leq n$ diagonal entries. Furthermore, we attain E by row and column reducing B.

We call the process of attaining the basis $\mathfrak{E}$ **normalizing** the matrix B.

More generally, a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$ given by a matrix A can be expressed as left multiplication by a normalized matrix $D = BAC$, where B is an $n \times n$ invertible matrix which applies column operations to A, and C is an $m \times m$ matrix performing row operations on A.

For the most part, we can apply these same ideas to integer matrices. But keep in mind $\mathbb{Z}^n$ does not quite have the structure of $\mathbb{R}^n$. One of the permissible matrix operations in $\mathrm{Mat}_{\mathbb{R}}(m, n)$ is scalar multiplication. A scalar is just a non-zero number in $\mathbb{R}$. But what characterizes a scalar is its existence of an inverse. In $\mathbb{Z}$ the only scalars, or **units**, are ± 1. Hence when we perform matrix operations we are only allowed to divide by ± 1.

Question

How does the normalization process change for integer matrices?

Suppose $\varphi : G \to H$ is a group homomorphism where $G \cong \mathbb{Z}^n$ and $H \cong \mathbb{Z}^m$. It would be very convenient to find bases $\mathfrak{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\} \subset G$ and $\mathfrak{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\} \subset H$ so that for some $r \leq n, m$,

$$\varphi(\mathbf{b}_i) = \begin{cases} \mathbf{c}_i & i \leq r \\ 0 & i > r \end{cases}.$$

Suppose φ is given by the $m \times n$ integer matrix A. Since $\mathbb{Z} \subset \mathbb{R}$, the following operations are permissible:

- add an integer multiple of one column (resp. row) to another
- interchange two columns (resp. rows)
- multiply a column (resp. row) by ± 1

Under these constraints, the desired matrix D will not be normalized, per se, but it will be diagonal. Theorem 1 gives an algorithm for producing a **canonical** $m \times n$ diagonal matrix D, given an $m \times n$ integer matrix A.

Smith normal form

Theorem 1 (Smith normal form)

Let A denote an $m \times n$ integer matrix. There exists an invertible $n \times n$ matrix B and an invertible $m \times m$ matrix C such that

$$D := BAC = \begin{pmatrix} d_1 & 0 & \cdots & & & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & d_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & \cdots & & & & 0 \end{pmatrix}$$

for some $r \leq n, m$ and $d_i | d_{i+1}$ for each $i = 1, \ldots, r-1$. In other words, each successive diagonal entry is an integer multiple of the last.

D is called the **Smith normal form** of A.

Proof.

We shall perform a sequence a row and column operations which culminate in the invertible matrices B, C. At each step we let $A = (a_{ij})$ denote the resulting matrix. Let R_i denote the ith row of A and let C_j denote the jth column.

Step 1: Locate the entry a with the smallest non-zero absolute value and permute rows and columns until $a_{11} = a$. If $a < 0$ then replace R_1 with $-R_1$.

Step 2: Clear the first column; given a a non-zero entry a_{i1}, write

$$a_{i1} = aq + r,$$

where q, r are as in the Division Algorithm (Equation (**??**)). In particular, $r < a$. Replace R_i with $R_i - qR_1$, and if $r \neq 0$, return to Step 1.

Repeating this process, each time we return to Step 1 the entry a gets strictly smaller. It will only take finitely many repetitions before either $a_{i1} = 0$ for all i or $a = 1$. But if $a = 1$ then Step 2 is required at most more time per entry to clear any remaining non-zero entries in the 1st column.

Clear the first row using an analogous process; given a non-zero entry a_{1j}, write

$$a_{1j} = aq + r$$

as in the Division Algorithm and replace C_j with $C_j - qC_1$. If $r \neq 0$ then return to Step 1. After finitely many repetitions a will be the only non-zero entry in the first row.

Step 3: Ensure the divisibilty condition; let B denote the submatrix of A given by omitting its first row and column. If an entry b of B, in the jth column of A, does not divide a, then replace C_1 with $C_1 + C_j$. The first column of A is no longer cleared so return to Step 2.

In Step 2 the Divison Algorithm will either replace b with zero or else direct us back to Step 1 where the entry a will become strictly smaller. Hence this process will end after finitely many steps. $\square$

The divisibility condition on the diagonal entries simply gives a canonical form for the matrix D and ensures its uniqueness.

The algorithm in the proof of Theorem 1 is more clear when seen in an example.

Example 2

Here, we apply the algorithm from Theorem 1 to a matrix A.

$$A := \begin{pmatrix} 1 & -1 & 1 \\ 5 & 1 & -5 \\ -3 & -3 & 29 \end{pmatrix} \xrightarrow[C_3 \to C_3 - C1]{C_2 \to C_2 + C_1} \begin{pmatrix} 1 & 0 & 0 \\ 5 & 6 & -10 \\ -3 & -6 & 32 \end{pmatrix}$$

$$\xrightarrow[R_3 \to R_3 - 5R_1]{R_2 \to R_2 - 5R_1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & -10 \\ 0 & -6 & 32 \end{pmatrix} \xrightarrow{C_3 \to C_3 + 2C_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & 2 \\ 0 & -6 & 20 \end{pmatrix}$$

$$\xrightarrow{C_2 \leftrightarrow C_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 6 \\ 0 & 20 & -6 \end{pmatrix} \xrightarrow{C_3 \to C_3 - 3C_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 20 & -66 \end{pmatrix}$$

$$\xrightarrow{R_3 \to -(R_3 - 10R_2)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 66 \end{pmatrix}.$$

Exercise 2 (cf. Problem 76)

Reduce the following matrices to Smith normal form.

(a) $\begin{pmatrix} 3 & 1 \\ -1 & 2 \end{pmatrix}$

(b) $\begin{pmatrix} 3 & 1 & -4 \\ 2 & -3 & 1 \\ -4 & 6 & -2 \end{pmatrix}$

Exercise 3 (cf. Problem 77)

Let Γ be the graph of Figure **??**. Reduce the transposed reduced Laplacian $\tilde{\Delta}$ of Γ.

Exercise 4 (cf. Problem 78)

Use the Sage command `smith_form` to diagonalize

$$\begin{pmatrix} 1 & 2 & 3 & -4 \\ -5 & 6 & 7 & 8 \\ -9 & -10 & 11 & 12 \\ 13 & 14 & -15 & 16 \end{pmatrix}.$$

What other information is obtained from `smith_form`?