MATH 3311, FALL 2025: LECTURE 35, NOVEMBER 21

Video: `https://youtu.be/-w5jGTFyPoM`
Smith Normal Form

Definition 1. A matrix $A \in M_{m\times n}(\mathbb{Z})$ is in **Smitn Normal Form** or **SNF** if:

(1) All its non-zero entrie are along the diagonal.
This means that the matrices
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{pmatrix}$$
are *not* in SNF.

(2) If a diagonal entry is 0 then all following diagonal entries are also 0.
This means that the matrices
$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
are not in SNF.

(3) If d_i is the i-th diagonal entry and d_{i+1} is the $(i+1)$-th entry, then $d_i \mid d_{i+1}$.
This means that the matrices
$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}, \begin{pmatrix} 4 & 0 & 0 \\ 0 & 6 & 0 \end{pmatrix}$$
are not in SNF.

(4) All the non-zero entries are *positive*.
This means that the matrix
$$\begin{pmatrix} -2 & 0 \\ 0 & 4 \end{pmatrix}$$
is not in SNF.

Theorem 1. *Given $A \in M_{m\times n}(\mathbb{Z})$, there exist $B \in \mathrm{GL}_m(\mathbb{Z})$ and $C \in \mathrm{GL}_n(\mathbb{Z})$ such that BAC is in Smith Normal Form.*

Remark 1. Even though the matrices B and C are *not* unique, the resulting SNF BAC *is* unique. Its entries can be determined as follows: An $r \times r$**-minor** of the matrix A is the determinant of an $r \times r$-submatrix (this can be any r rows and r columns picked out of the rows and columns of A). For instance, a 1×1-minor is just an entry of the matrix. Now, if $d_1, d_2, \ldots, d_s$ are the non-zero entries of the SNF, then $d_1 d_2 \cdots d_r$ is the gcd of the $r \times r$-minors.

Caution 1. Do *not* use this to compute the SNF! This is computationally not viable beyond the 2×2-case.

Example 1. Let us look at the matrix
$$\begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix}$$
Given that we know an SNF exists, we can actually figure out what it is by 'pure thought'. The key point is that multiplying by invertible integer matrices doesn't affect the determinant up to sign. Therefore, the SNF will have to have determinant 24. The possibilities are
$$\begin{pmatrix} 1 & 0 \\ 0 & 24 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 12 \end{pmatrix}$$
Now, if we multiply our given matrix by another integer matrix, we can only get even entries (why?). Therefore, there is no way to get 1 via such a process. This means that the second guy must be the SNF.

To actually see that this is the SNF concretely, we want to see that we can multiply the matrix on the left and the right by invertible matrices to get to the supposed SNF. As we saw last time, we can interpret some of these multiplication operations concretely in terms of one of the following operations:

(i) Adding an integer multiple of one row to a *different* row.
(i') Adding an integer multiple of one column to a *different* column.
(ii) Switching two rows.
(ii') Switching two columns.

We can apply these operations to our 2×2-matrix:

$$\begin{aligned}
\begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix} &\xrightarrow{C1 \mapsto C1+C2} \begin{pmatrix} 4 & 0 \\ 6 & 6 \end{pmatrix} \\
&\xrightarrow{R2 \mapsto R2-R1} \begin{pmatrix} 4 & 0 \\ 2 & 6 \end{pmatrix} \\
&\xrightarrow{R1 \leftrightarrow R2} \begin{pmatrix} 2 & 6 \\ 4 & 0 \end{pmatrix} \\
&\xrightarrow{R2 \mapsto R2-2 \cdot R1} \begin{pmatrix} 2 & 6 \\ 0 & -12 \end{pmatrix} \\
&\xrightarrow{C2 \mapsto C2-3 \cdot C1} \begin{pmatrix} 2 & 0 \\ 0 & -12 \end{pmatrix}
\end{aligned}$$

At this point, we want to get rid of the minus sign in the bottom right. For this, we see that the following operations are also permissible:

(iii) Multiplying a row by -1;
(iii') Multiplying a column by -1.

Applying this to the second row or second column now changes the entry to 12 and gives us the SNF.

Remark 2. What happened here is that once we got the 2, which is the gcd of all the entries into the top left corner, we could use it to get rid of other entries in the first row and column.

Example 2. Consider instead the matrix

$$\begin{pmatrix} 16 & 0 \\ 0 & 6 \end{pmatrix}$$

We can perform the following operations right away:

$$\begin{pmatrix} 16 & 0 \\ 0 & 6 \end{pmatrix} \xrightarrow{C1 \mapsto C1+C2} \begin{pmatrix} 16 & 0 \\ 6 & 6 \end{pmatrix}$$

Here, unlike in the previous case, we cannot immediately bring $\gcd(16, 6) = 2$ into the top left by doing the operations we've seen already. It is not impossible, but is a little tedious. One can circumvent this by using Bezout! Namely, we can get 2 as a linear combination

$$2 \cdot 16 + (-5) \cdot 6 = 2.$$

Therefore, if we were somehow able to use a matrix multiplication to convert $R1$ to $2 \cdot R1 + (-5) \cdot R2$, then we would be in good shape. A bit of thought shows that we need to find an *invertible* matrix with first row entries given by $2, -5$. Since

$$2 \cdot 8 + (-5) \cdot 3 = 1,$$

we find that the matrix

$$\begin{pmatrix} 2 & -5 \\ -3 & 8 \end{pmatrix}$$

is invertible: its determinant is 1. Moreover, we have

$$\begin{pmatrix} 2 & -5 \\ -3 & 8 \end{pmatrix} \cdot \begin{pmatrix} 16 & 0 \\ 6 & 6 \end{pmatrix} = \begin{pmatrix} 2 & -30 \\ 0 & 48 \end{pmatrix}$$

At this point, I can add 15 times C1 to C2 to get rid of the -30, leaving me with the SNF

$$\begin{pmatrix} 2 & 0 \\ 0 & 48 \end{pmatrix}$$

Proof of Theorem 1. The basic idea of finding the SNF is to get the gcd of the entries into the top left corner, use it to get rid of the other entries in the first row and column, and to repeat the process with the remaining matrix consisting of the other rows and columns.

(i) If A is the zero matrix, we are done.
(ii) Otherwise, our first goal is to get the gcd of all the non-zero entries in A into the top left corner.
(iii) If this gcd is one of the entries of A, then doing row and column swaps will get that entry to the top left entry.
(iv) Otherwise, ensure that the top left entry is non-zero, and then, by repeatedly applying operations (ii),(ii') and using invertible matrices defined using Bezout as in the previous example, get to the point where the top left entry a_0 is the gcd of all the non-zero entries in the first row and column.
(v) Now, every entry in the first row and column is an integer multiple of a_0. Therefore, by repeatedly applying operations (i) and (i'), and subtracting appropriate multiples of the first column and row from the others, we can ensure that a_0 is the only *non-zero* entry both in the first row *and* in the first column.
(vi) If a_0 is now the gcd of all the non-zero entries of the matrix, then we are done. Otherwise, we find the first non-zero entry a_1 that a_0 does not divide, and add the corresponding row to the first row. We repeat the previous two steps to replace a_0 in the top left entry with the gcd of a_0 and a_1, while still ensuring that it is the only non-zero entry in the first row and column.
(vii) We now repeat the last step for each other non-zero entry till the top left entry (still denoted a_0) is the gcd of all the entries of the matrix.
(viii) We have now arrived at a block matrix of the form

$$\begin{bmatrix} a_0 & 0 \\ 0 & A_1 \end{bmatrix},$$

where a_0 is the gcd of all the non-zero entries and A_1 is an $(r-1)\times(s-1)$-matrix. Inductively applying steps (i)-(viii) to A_1 will now finish the process of finding the Smith Normal Form.

□

Back to finitely generated abelian groups

Let us recap how we aim to prove the existence part of the Fundamental Theorem of finitely generated abelian groups:

(1) First, given such a group G, we can choose a set of generators $\{x_1, \ldots, x_m\} \subset G$ and use this to write down a *surjective* homomorphism $\mathbb{Z}^m \to G$ carrying $\overrightarrow{\mathbf{e}}_i$ to x_i.
(2) Such a homomorphism gives an isomorphism $\mathbb{Z}^m/H \xrightarrow{\simeq} G$ for some subgroup $H \leq \mathbb{Z}^m$.
(3) The subgroup H is isomorphic to $\mathbb{Z}^n$ for some $n \geq m$, and so we can find n vectors $\overrightarrow{\mathbf{v}}_1, \ldots, \overrightarrow{\mathbf{v}}_n \in H$ that generate H. More conceptually, we have a homomorphism $\alpha : \mathbb{Z}^n \to \mathbb{Z}^m$ whose image is H.
(4) Write down the $m \times n$-matrix

$$A_\alpha = \begin{pmatrix} \overrightarrow{\mathbf{v}}_1 & \cdots & \overrightarrow{\mathbf{v}}_n \end{pmatrix}$$

(5) Find the Smith Normal Form for A_α: this is of the form $A_\psi A_\alpha A_\varphi$, where A_ψ and A_φ are invertible matrices associated with *isomorphisms* $\psi : \mathbb{Z}^m \xrightarrow{\simeq} \mathbb{Z}^m$ and $\varphi : \mathbb{Z}^n \to \mathbb{Z}^n$.
(6) If $d_1 \mid d_2 \mid \cdots \mid d_n$ are the diagonal entries of $A_\psi A_\alpha A_\varphi$, then see that

$$\mathbb{Z}^m/\mathrm{im}\,(\psi \circ \alpha \circ \varphi) \simeq \mathbb{Z}/d_1\mathbb{Z} \times \cdots \times \mathbb{Z}/d_n\mathbb{Z} \times \mathbb{Z}^{m-n}.$$

(7) Observe that this has to be isomorphic to $\mathbb{Z}^m/\mathrm{im}\,\alpha = \mathbb{Z}^m/H \xrightarrow{\simeq} G$.