Week 4, September 20

Seewoo Lee

1 Discussion notes

- Linear map: We define the notion of injectivity, surjectivity, and bijectivity of linear maps. We just apply the usual definition of them for "functions". But since we are considering linear maps, not just random functions, we can describe them in different ways: for $T : V \to W$ linear,

  - T is injective if and only if $\mathrm{null}(T) = \{0\}$.
  - T is surjective if and only if $\mathrm{range}(T) = W$.

  Of course, T is bijective if and only if T is both injective and surjective.

  One of the most important theorem in linear algebra is the Fundamental Theorem of Linear Maps (FTLM), which relates dimension of the null space and the range (3.21):

  $$\dim \mathrm{null}(T) + \dim \mathrm{range}(T) = \dim V.$$

  The proof goes as follows: choose a basis $\{u_1, \dots, u_n\}$ of $\mathrm{null}(T)$, and extend it to a basis of V, by adding more vectors $\{v_1, \dots, v_m\}$. Then one can prove that $\{Tv_1, \dots, Tv_m\}$ form a basis of $\mathrm{range}(T)$. It has a lot of applications. For example, when $\dim V > \dim W$, one can prove that $T : V \to W$ is never injective (3.22), which can be considered a "pigeonhole principle for linear maps" (there are $\dim V$ many pigeons, and we want to put them (apply T) into $\dim W$ many holes).[1] Similarly, when $\dim V < \dim W$, there are no surjective linear maps from V to W (3.24, there are more pigeonholes than pigeons, so some of them would be empty). See "Additional Problem 2" below to see what happens if $\dim V = \dim W$.

[1] Note from a discussion: this is not true for general maps that are not linear. In fact, there exists a bijection between $\mathbb{R}^5$ and $\mathbb{R}^3$. It is not easy to describe explicitly, but you can check this MO question and answer for the bijection between $\mathbb{R}$ and $\mathbb{R}^2$, and you can use it to build a bijection between $\mathbb{R}^5$ and $\mathbb{R}^3$.

- Matrix representation: We can summarize 3C (or whole Chapter 3) as follows:

$$\text{linear map} + \text{choice of basis} \leftrightarrow \text{matrix}.$$

For $\to$ direction, the corresponding matrix is defined as follows (3.31): let $\beta = \{v_1, \ldots, v_n\}$ and $\gamma = \{w_1, \ldots, w_m\}$ be basis of V and W, respectively. We have

$$\begin{aligned} Tv_1 &= a_{1,1}w_1 + a_{2,1}w_2 + \cdots + a_{m,1}w_m \\ Tv_2 &= a_{1,2}w_1 + a_{2,2}w_2 + \cdots + a_{m,2}w_m \\ &\vdots \\ Tv_n &= a_{1,n}w_1 + a_{2,n}w_2 + \cdots + a_{m,n}w_m \end{aligned}$$

and the corresponding $m \times n$ matrix

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & & & \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

and we denote the matrix ax $M(T, \beta, \gamma)$, or $M(T)^{\gamma}_{\beta}$, or $[T]^{\gamma}_{\beta}$. When $T : V \to V$, we simply write $[T]_{\beta}$ (using *same* β for V and V). Note that you need to somehow "transpose" the array of numbers you get from the previous equations. This essentially happens because we are considering matrix representations as "multiplying a matrix to the left of a column vector", not "multiplying a matrix to the right of a row vector".

This association gives a map $\mathcal{L}(V, W)$ to $\mathbf{F}^{m,n}$, and it is even a linear map: $M(T_1 + T_2) = M(T_1) + M(T_2)$ and $M(cT) = cM(T)$. More importantly, composition of linear maps correspond to the multiplication of matrices: for

$$V \xrightarrow{T} W \xrightarrow{S} U$$

choose basis β, γ, δ for V, W, U respectively[2] and we get associated matrices $[T]^{\gamma}_{\beta}$, $[S]^{\delta}_{\gamma}$. The composition $ST : V \to U$ is also a linear map, and we have a corresponding matrix representation $[ST]^{\delta}_{\beta}$. Then we have:

$$[ST]^{\delta}_{\beta} = [S]^{\delta}_{\gamma} \cdot [T]^{\gamma}_{\beta}$$

[2]Why not α? Don't ask...

Under our philosophy on linear maps and matrices, we can summarize these as follows: addition, scalar multiplication, and matrix multiplications are *defined* in view of addition / scalar multiplication / composition of linear maps. I strongly suggest you to write down the multiplication and composition part explicitly when $\dim V = \dim W = \dim U = 2$, but only once in your life.

- Rank of a matrix (3.52): For a given matrix, we define *column* rank and *row* rank, as dimensions of the column space and the row space, respectively. We have a column-row factorization (3.56): any rank c m-by-n matrix can be decomposed as $A = CR$ where $C \in \mathbf{F}^{m,c}$ and $R \in \mathbf{F}^{c,n}$. (See what happens in the extreme case of $c = 1$). Using this, we can prove that column rank is equal to the row rank (3.57), and we call it as a *rank*.

  We can also define a rank of a linear map, but you need to wait until Chapter 3D: we don't know *a priori* whether a rank of a matrix representation $[T]_\beta^\gamma$ depends on a choice of basis β, γ or not. You'll eventually find that the rank *does not* depend on a choice of basis, so that we can define a rank of a linear map, without referring to a basis. You can also say that the rank of T is equal to the dimension of the range(T).

2 Problems

- Recommended problems: 3B.12, 3B.15 (without using FTLM - can you construct a basis of V using a basis of null(V) and range(V)?), 3B.23, 3B.27, 3B.29, 3C.11, 3C.12.

- Additional problems:

  1. Prove that linear map sends an additive inverse to an additive inverse: if $Tv = w$, then $T(-v) = -w$.

  2. Let $T : V \to W$ be a linear map and $\dim V = \dim W$. Prove that T is injective if and only if T is surjective.

  3. (Bad exercise) Try to solve 3B.14 without using FTLM, but only with matrices. This motivates you to study something "abstract but simpler" things.

  4. (Some machine learning?) Theorem 3.56 has a practical application in machine learning, so-called *parameter-efficient fine-tuning*.

     (a) Let M be a $m \times n$ matrix over $\mathbb{R}$. How many numbers in M?

(b) Assuming that you're multiplying M with a column vector $v \in \mathbb{R}^{n,1}$. How many *multiplications* you need in the computation?

(c) Now, assume that M has rank c, so that it can be written as $M = CR$ where $C \in \mathbb{R}^{m,c}$ and $R \in \mathbb{R}^{c,n}$. How many numbers do you need to represent M? If you multiply CR to $v \in \mathbb{R}^{n,1}$, how many multiplications do you need? You can compute the numbers with $m = 100$, $n = 200$, and $c = 5$, and compare the numbers from (a) and (b).

This can be used to train ML models more efficiently, and make them "smaller". If you are interested in these things, check our the paper "LoRA: Low-Rank Adaptation of Large Language Models". (You can also check this blog post from a random person...)