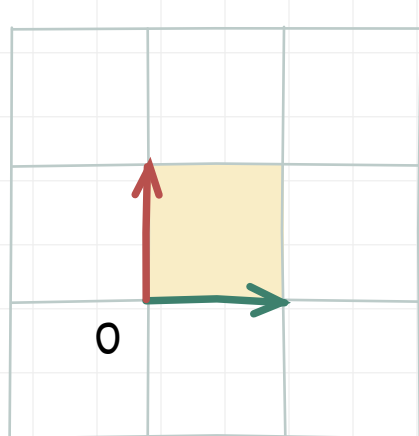


Ma trận = làm gì với không gian?

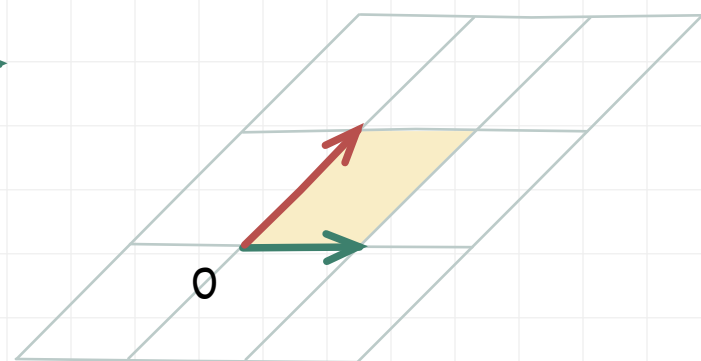
Tưởng tượng kéo một tấm lưới: hàng càng cao, trượt càng xa.

trước

sau: trượt ngang



một ô vuông



một hình bình hành

Nhìn hình để viết tọa độ

Điểm cao y thì trượt sang phải y .

Chiều cao giữ nguyên: (x, y) thành $(x + y, y)$.

$$S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad S(x, y) = (x + y, y)$$

"Tuyến tính" giữ cách ghép mũi tên

$$M(u + v) = Mu + Mv$$

$$M(cu) = c(Mu)$$

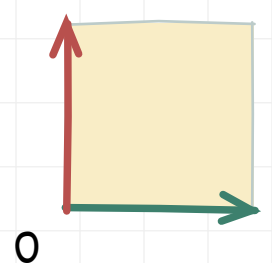
Ghép rồi biến đổi = biến đổi từng mũi tên rồi ghép.

Gốc vẫn ở gốc; đường thẳng có thể bị ép thành một điểm.

Hai mũi tên quyết định tất cả

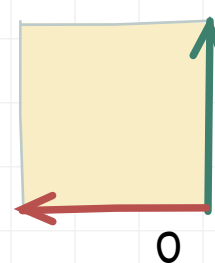
Quay cả mặt phẳng 90° ngược chiều kim đồng hồ.

$$e_2 = (0, 1)$$



$$e_1 = (1, 0)$$

quay $\rightarrow$



$$Re_1 = (0, 1)$$

$$Re_2 = (-1, 0)$$

Ghi nơi đáp xuống vào từng cột

$$R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

— cột 1: ảnh của e_1
— cột 2: ảnh của e_2

Vì mọi vector đều ghép từ e_1, e_2

$$v = (3, 2) = 3e_1 + 2e_2$$



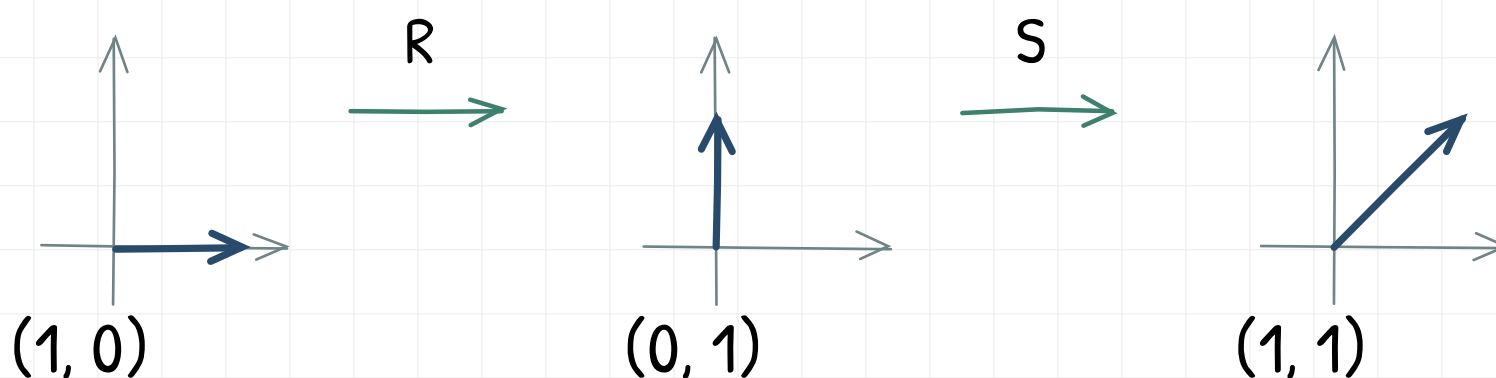
$$\begin{aligned} Rv &= 3Re_1 + 2Re_2 \\ &= 3(0, 1) + 2(-1, 0) = (-2, 3) \end{aligned}$$

$$Mv = x \times \text{cột 1} + y \times \text{cột 2}$$

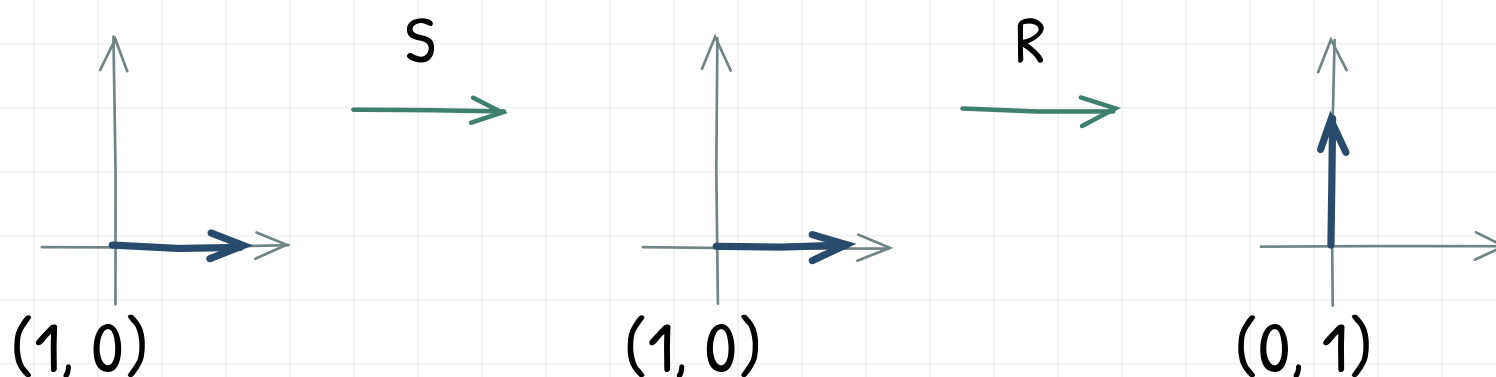
Đổi thứ tự, đổi nơi đáp xuống

R: quay 90° S: trượt ngang Bắt đầu ở $e_1 = (1, 0)$.

Quay rồi trượt



Trượt rồi quay



SR nghĩa là R trước, S sau

$$SR = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$$

$$RS = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$$

Hai hành động nói chung không đổi chỗ được: $AB \neq BA$.

Nhưng $(AB)C = A(BC)$: vẫn C rồi B rồi A.

Chỉ đổi cách gộp; thứ tự thao tác không đổi.

Hàng $\times$ cột từ đâu mà ra?

Theo một mũi tên qua B rồi A để tìm cột của AB.

$$e_2 \xrightarrow{B} (6, 8) \xrightarrow{A} (22, 50)$$

$$\begin{matrix} & A \\ \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} & \times \end{matrix} \begin{matrix} & B \\ \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} & = \end{matrix} \begin{matrix} & AB \\ \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} \end{matrix}$$

Nhìn riêng cột 2:

$$\begin{aligned} A(6, 8) &= 6 \times (1, 3) + 8 \times (2, 4) \\ &= (22, 50) \end{aligned}$$

Lấy tọa độ đầu: hàng 1 chấm cột 2

$$22 = (1, 2) \cdot (6, 8) = 1 \times 6 + 2 \times 8$$

Mỗi ô là một tích vô hướng (bài 1).

$$(AB)[i, j] = \text{tổng theo } t \text{ của } A[i, t] \times B[t, j]$$

Kích thước = đầu ra nối vừa đầu vào

$$(m \times n) \times (n \times p) = m \times p$$

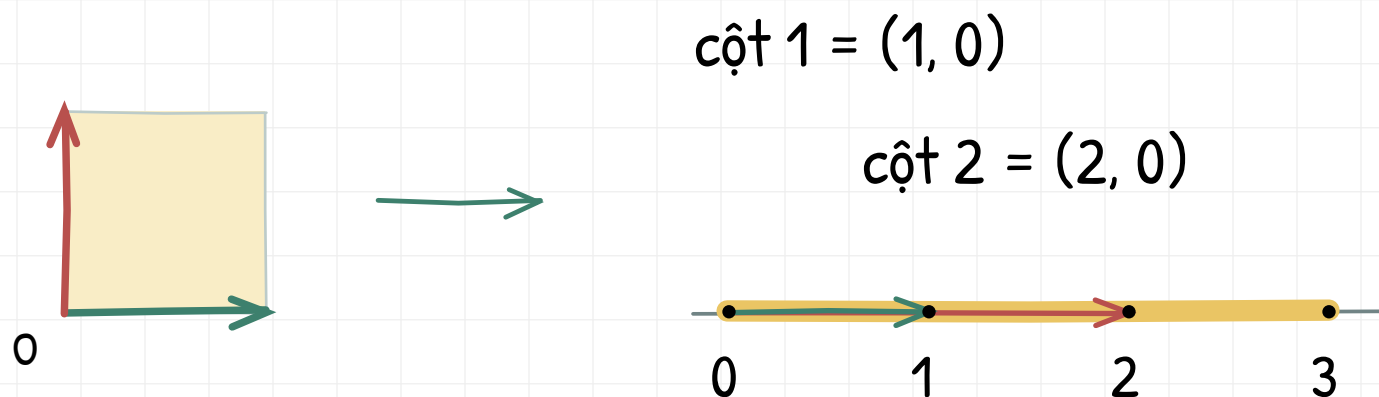
B: p số vào $\rightarrow$ n số ra; A nhận n số $\rightarrow$ trả m số.

Hai số giữa phải khớp. Kết quả có hai số ngoài.

☆ NumPy: $A @ B$ là nhân ma trận; $A * B$ là nhân từng ô.

Bóp dẹp: thông tin đi đâu?

Ép cả ô vuông xuống một đường thẳng.



Hai cột cùng phương

$$P = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

$$P(x, y) = (x + 2y, 0)$$

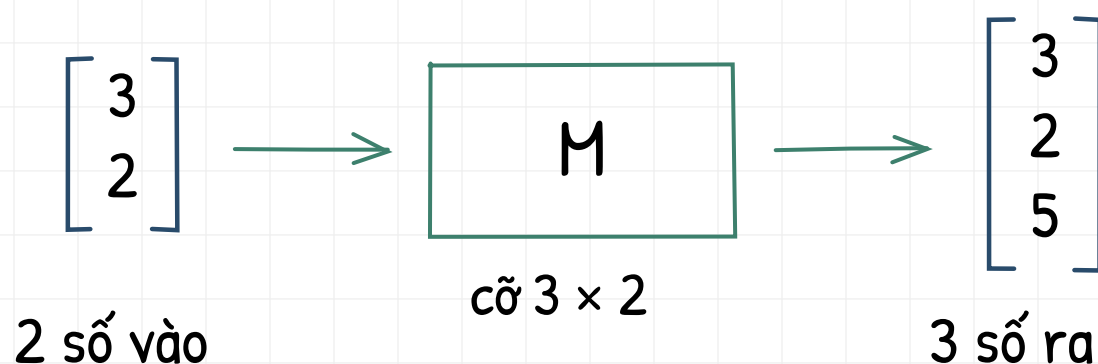
Mọi đầu ra đều nằm trên trục ngang.

Mọi tổ hợp hai cột vẫn ở trên một đường.

Chỉ còn 1 hướng độc lập: hạng = 1 (bài 3).

(2, 0) và (0, 1) đều thành (2, 0): không phân biệt lại được.

Ma trận không nhất thiết vuông

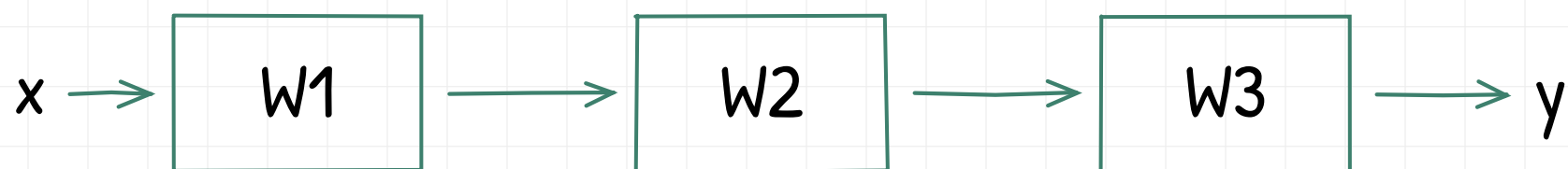


M có 3 hàng (1, 0), (0, 1), (1, 1).

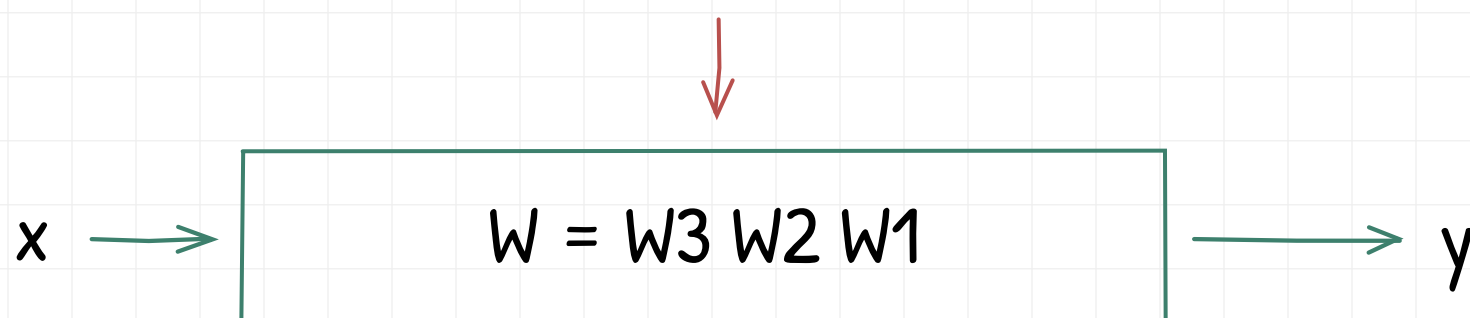
Tổng quát: $m \times n$ biến n tọa độ thành m tọa độ.

Vì sao mạng cần phi tuyến?

Ba lần kéo / quay / trượt vẫn là một phép tuyến tính.

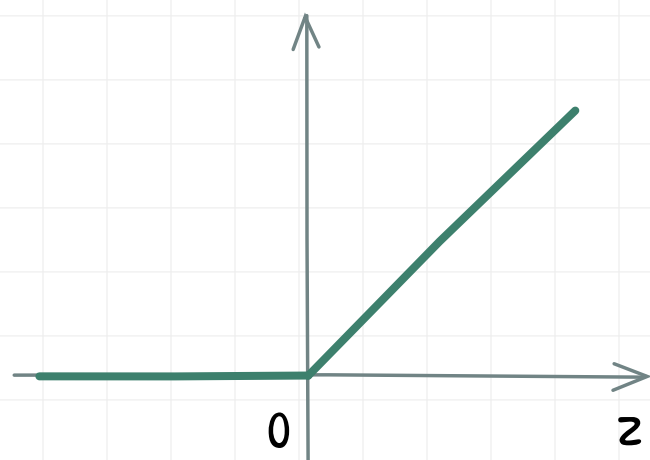


gộp cả chuỗi lại



$$W3(W2(W1x)) = (W3W2W1)x$$

Chèn một “góc gãy” vào giữa



$$\text{ReLU}(z) = \max(0, z)$$

Phía âm bị ép về 0.

Phía dương giữ nguyên.

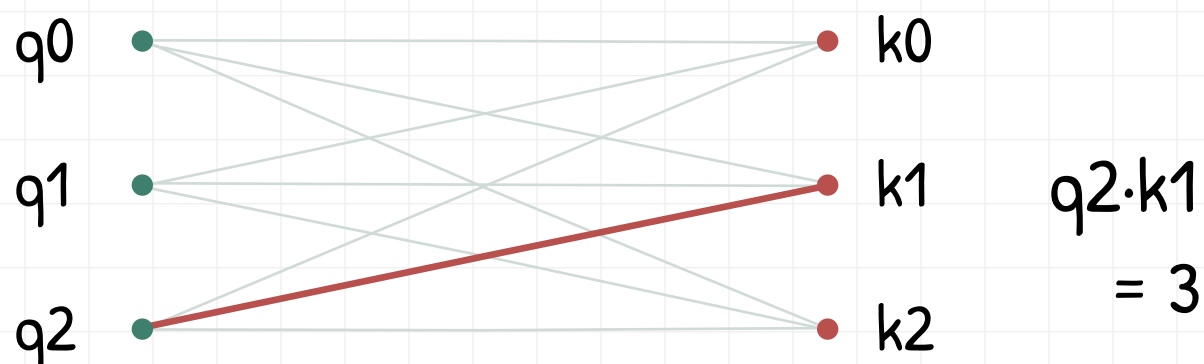
Đường bị gãy: không thể là một ma trận duy nhất.

$W2 \text{ ReLU}(W1x)$: nói chung không gộp thành Wx .

Phép nhân vẫn kết hợp; hàm phi tuyến chặn việc gộp ma trận.

Attention: mỗi ô là một cặp

Mỗi truy vấn gặp từng khóa: $3 \times 3 = 9$ phép đo.



Mỗi hàng là một vector; chuyển vị K để hàng thành cột.

$$\begin{array}{ccc}
 Q & K & QK^T \\
 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} & \begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 1 & 1 \end{bmatrix} & \begin{bmatrix} 2 & 0 & 1 \\ 0 & 3 & 1 \\ 2 & 3 & 2 \end{bmatrix}
 \end{array}$$

$$(3 \times 2) \times (2 \times 3) = 3 \times 3$$

Ô $[2, 1] = (1, 1) \cdot (0, 3) = 3$. Chỉ số tính từ 0.

Đây là điểm dot: có cả hướng lẫn độ dài, chưa phải xác suất.

Cùng kết quả, khác chi phí

$$N \times N$$

$$(QK^T)V$$

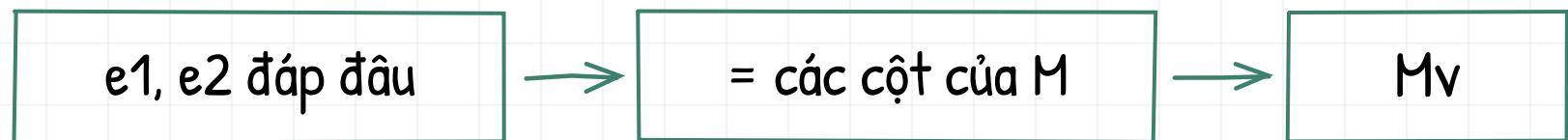
$$d \times d$$

$$Q(K^TV)$$

Q, K, V cùng cỡ $N \times d$; N lớn hơn d : cách phải thường rẻ hơn.

Chỉ đổi cách gộp khi không có softmax chen giữa.

Tổng kết & bảng công thức



$$Mv = x \times \text{cột 1} + y \times \text{cột 2} \quad ABv = A(Bv): B \text{ trước, } A \text{ sau}$$

Bảng công thức

$A + B$: cộng từng ô	$(A + B)v = Av + Bv$
cA : nhân từng ô với c	$(cA)v = c(Av)$: co giãn kết quả
$(AB)C = A(BC)$	luôn đúng: chỉ đổi cách gộp
$A(B + C) = AB + AC$	luôn đúng
$AB \neq BA$ nói chung	$IA = AI = A$: I không làm gì
$A^T[i, j] = A[j, i]$	lật qua đường chéo; $m \times n \rightarrow n \times m$
$(A^T)^T = A, a \cdot b = a^T b$	dot = nhân $(1 \times n)(n \times 1)$
$R^T = R^{-1}$ (phép quay)	chuyển vị = quay ngược lại
$(AB)^T = B^T A^T$	ĐẢO thứ tự
$A^{-1}A = I, (AB)^{-1} = B^{-1}A^{-1}$	ĐẢO thứ tự

☆ Gỡ một chuỗi hành động: gỡ cái làm SAU CÙNG trước.
Nhu mặc tất rồi giày, thì cởi giày rồi mới cởi tất.

Chuyển vị không phải nghịch đảo; chỉ trùng nhau ở phép quay.
Nghịch đảo 2×2 và khi nào không có A^{-1} : bài 4.

Tự kiểm, rồi mới sang bài 3

1. Vì sao chỉ cần biết e_1, e_2 đáp đầu là biết trọn M ?
2. Vì sao $AB \neq BA$? Giải thích bằng hành động.
3. Vậy sao $(AB)C = A(BC)$ thì luôn đúng?
4. I làm gì? Vì sao $IA = AI = A$?
5. M ép mặt phẳng thành một đường: hai cột ra sao?
6. Vì sao 3 lớp tuyến tính = 1 lớp? Suy ra cần gì?
7. Ma trận 3×2 đi từ mấy chiều sang mấy chiều? 2×3 ?
8. Q, K có 3 hàng 2 chiều: QK^T cỡ nào, ô (i, j) đo gì?

Đáp nhanh

1. $v = x e_1 + y e_2$, M giữ phép cộng và nhân số
 $\Rightarrow Mv = x \times \text{cột } 1 + y \times \text{cột } 2$.
2. Phép sau tác động lên kết quả của phép trước;
 đổi thứ tự là đổi đầu vào của từng phép.
3. Cả hai đều làm C , rồi B , rồi A ; chỉ đổi cách gộp.
4. I không làm gì: đứng trước hay sau A đều chỉ còn A .
5. Cùng phương (cột này là bội cột kia): hạng 1.
6. $W_3 W_2 W_1$ là một ma trận $\Rightarrow$ cần phi tuyến chen giữa.
7. 3×2 : 2 chiều $\rightarrow$ 3 chiều. 2×3 : 3 chiều $\rightarrow$ 2 chiều.
8. 3×3 ; ô $(i, j) = q_i \cdot k_j$: dot truy vấn i với khóa j .