VII.1 The Construction of Deep Neural Networks

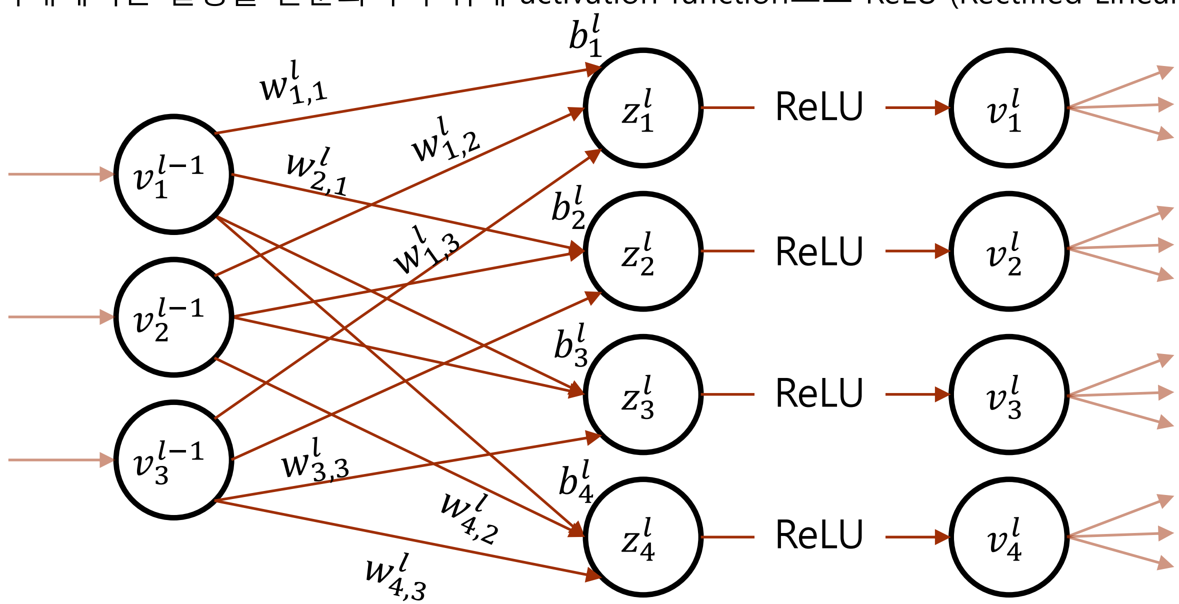
- 신경망, 선형방정식, 최적화
  - 3가지 주제들의 관계 및 앞으로 사용할 notation 요약
- 신경망 문제의 일반화
  - 예) 손글씨 숫자의 인식 (MNIST)
    - 각각의 샘플 입력 데이터는 28x28개의 픽셀에 대한 grey scale값으로 이루어짐 ➔ 784(=28x28)차원 공간($\mathbf{R}^{784}$)내의 벡터로 표현 가능
    - 출력 값은 0부터 9까지 숫자 각각에 해당되는 확률의 배열 ➔ 확률들을 성분으로 가진 10차원 공간($\mathbf{R}^{10}$) 내의 벡터 형태로 표현 가능 $(p_0, p_1, \cdots, p_9)$
  - 각각의 training 데이터 샘플들을 N개의 특징(feature)를 가진 입력 벡터$\boldsymbol{v}$와 M개의 값을 가진 출력 벡터 $\boldsymbol{y}$로 일반화하면, $\boldsymbol{y} = \boldsymbol{F}_L\left(\boldsymbol{F}_{L-1}\left(\cdots\left(\boldsymbol{F}_1(\boldsymbol{v})\right)\right)\right) = \boldsymbol{F}(\boldsymbol{v})$의 관계에 가까운 $\boldsymbol{F}$를 찾는 문제임

VII.1 The Construction of Deep Neural Networks

- (계속) 신경망 문제의 일반화
  - $\boldsymbol{y} = \boldsymbol{F}_L\left(\boldsymbol{F}_{L-1}\left(\cdots\left(\boldsymbol{F}_1(\boldsymbol{v})\right)\right)\right) = \boldsymbol{F}(\boldsymbol{v})$
  - 신경망의 구현에는 위의 $\boldsymbol{F}_l$와 같이 l-번째 인접한 layer들 간의 계산시 다음과 같은 관계 존재
    - 인접한 layer의 값 계산에는 부분적으로 선형적인 관계가 존재.

$$z_j^l = \sum_i w_{j,i}^l v_i^{l-1} + b_j^l \quad \Leftrightarrow \quad \boldsymbol{z}^l = W^l \boldsymbol{v}^{l-1} + \boldsymbol{b}^l$$

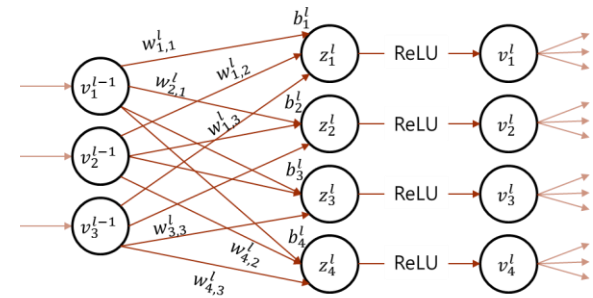
    - $w_{j,i}^l$는 $(l-1)$-번째 layer의 i-번째 neuron의 출력이 l-번째 layer의 j-번째 neuron의 입력에 영향을 끼치는 비중 (weight)
    - b_j^l는 activation function 전에 더해지는 bias 값
    - 아래에서는 설명을 단순화하기 위해 activation function으로 ReLU (Rectified Linear Unit) 고려



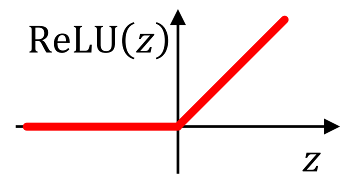
VII.1 The Construction of Deep Neural Networks

- (계속) 신경망 문제의 일반화

$$z_j^l = \sum_i w_{j,i}^l v_i^{l-1} + b_j^l \quad \Leftrightarrow \quad \boldsymbol{z}^l = W^l \boldsymbol{v}^{l-1} + \boldsymbol{b}^l$$



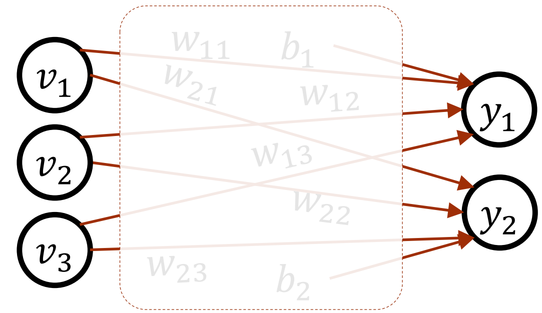
  - ReLU (Rectified Linear Unit)
    - $\mathrm{ReLU}(z) = \max(z, 0) = \begin{cases} 0, z < 0 \\ z, z \geq 0 \end{cases}$
    - z는 다른 변수들의 1차결합 형태이므로 연속이면서 부분적으로 선형인 함수 (CPL: continuous piecewise linear function)의 형태를 띔
    - 일반적으로 layer의 수와 같은 layer내에 있는 neuron의 개수가 증가함에 따라 같은 선형 상수를 사용할 수 있는 구간은 좁아짐
    - 좁은 영역의 입력에 대해서는 입력과 출력 간의 선형 관계를 고려할 수 있음



  - 다른 activation 함수들과의 비교
    - 완전한 선형 함수는 일반적으로 도움이 안됨
    - 실제 neuron과 유사한 step function 은 미분이 0이라 backpropagation 을 이용한 최적화가 어려움
    - ReLU와 유사한 sigmoid 함수$\left(\frac{1}{1+e^{-x}}\right)$도 큰 값에서 미분이 0이 되는 단점이 존재

VII.1 The Construction of Deep Neural Networks

- (계속) 신경망 문제의 일반화
  - 최적화의 기본 개념들의 설명을 위해 좁은 입력 범위에 대해 전체 함수 $\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{v})$가 선형 관계로 표현되었다고 가정



- Example
  - Training dataset에 3개의 샘플 데이터가 있다고 가정
  - i-번째 샘플 데이터는 3개의 입력 feature $(v_{i,1}, v_{i,2}, v_{i,3})$와 2개의 출력값 $(y_{i,1}, y_{i,2})$을 가진다고 가정

첫번째 샘플 데이터

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix}$$

두번째 샘플 데이터

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \\ v_{23} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix}$$

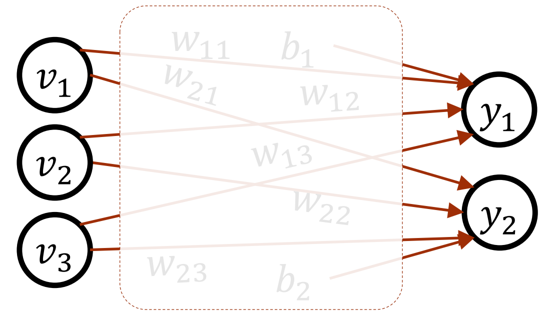
세번째 샘플 데이터

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{31} \\ v_{32} \\ v_{33} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{31} \\ y_{32} \end{bmatrix}$$

$$w_{11}v_{11} + w_{12}v_{12} + w_{13}v_{13} + b_1 = y_{11}$$
$$w_{21}v_{11} + w_{22}v_{12} + w_{23}v_{13} + b_2 = y_{12}$$
$$w_{11}v_{21} + w_{12}v_{22} + w_{13}v_{23} + b_1 = y_{21}$$
$$w_{21}v_{21} + w_{22}v_{22} + w_{23}v_{23} + b_2 = y_{22}$$
$$w_{11}v_{31} + w_{12}v_{32} + w_{13}v_{33} + b_1 = y_{31}$$
$$w_{21}v_{31} + w_{22}v_{32} + w_{23}v_{33} + b_2 = y_{32}$$

VII.1 The Construction of Deep Neural Networks

- Example
  - Training dataset에 3개의 샘플 데이터가 있다고 가정
  - i-번째 샘플 데이터는 3개의 입력 feature $(v_{i,1}, v_{i,2}, v_{i,3})$와 2개의 출력값 $(y_{i,1}, y_{i,2})$을 가진다고 가정



$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix}$$

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \\ v_{23} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix}$$

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} v_{31} \\ v_{32} \\ v_{33} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{31} \\ y_{32} \end{bmatrix}$$

입력 출력

$$\begin{bmatrix} v_{11} & v_{12} & v_{13} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & v_{11} & v_{12} & v_{13} & 1 \\ v_{11} & v_{12} & v_{13} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & v_{11} & v_{12} & v_{13} & 1 \\ v_{11} & v_{12} & v_{13} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & v_{11} & v_{12} & v_{13} & 1 \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{12} \\ w_{13} \\ b_1 \\ w_{21} \\ w_{22} \\ w_{23} \\ b_2 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}$$

최적화의 대상

$\boldsymbol{Ax} = \boldsymbol{b}$ 형태의 문제를 풀어야 함

$W\boldsymbol{v_i} + \boldsymbol{b} = \boldsymbol{y}$의 형태

$\Leftrightarrow W^l \boldsymbol{v}^{l-1} + \boldsymbol{b}^l = \boldsymbol{z}^l$ 와 유사

II.2 Least Squares: Four Ways

- $A\boldsymbol{x} = \boldsymbol{b}$의 해를 구하는 문제
  - 만약 행렬 A의 크기가 $n \times n$인 정사각 행렬이고, rank r도 $r = n$ 이면 항상 1개의 해가 존재
  - 행렬 A의 크기가 일반적인 $m \times n$인 직사각형일 경우, rank r이 $r = m$를 만족하면 해는 1개 또는 여러 개가 존재
  - $r < m$인 경우에는 벡터 $\boldsymbol{b}$가 행렬 A의 열벡터 공간에 포함되는 경우에는 해가 존재하고, 포함되지 않으면 정확한 해는 존재하지 않음
- 최소 자승법 (Least squares)
  - $\|\boldsymbol{b} - A\boldsymbol{x}\|^2$를 최소로 만드는 $\boldsymbol{x} = \hat{\boldsymbol{x}}$해를 찾는 방법
  - $A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b}$
  - 유사 역행렬 (pseudoinverse) A^+이용
  - $A = QR$
  - $\|\boldsymbol{b} - A\hat{\boldsymbol{x}}\|^2 + \delta^2\|\hat{\boldsymbol{x}}\|^2$를 최소화

II.2 The Normal Equations $A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b}$

- $\|\boldsymbol{b} - A\boldsymbol{x}\|^2$의 최소화
  - $L(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$ 는 일반적으로 loss function 또는 cost function이라고 불림

    $\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = (\boldsymbol{b} - A\boldsymbol{x})^T(\boldsymbol{b} - A\boldsymbol{x}) = (\boldsymbol{b}^T - \boldsymbol{x}^T A^T)(\boldsymbol{b} - A\boldsymbol{x})$

    $= \boldsymbol{x}^T A^T A\boldsymbol{x} - \boldsymbol{b}^T A\boldsymbol{x} - \boldsymbol{x}^T A^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b}$
  - $L(\boldsymbol{x})$ 가 $\boldsymbol{x} = \hat{\boldsymbol{x}}$에서 최소값을 가지면, 변수들로 이루어진 벡터 $\boldsymbol{x}^T = [x_1 \quad \cdots \quad x_n]$ 의 모든 성분 변수에 대해 다음의 관계가 성립되어야 함.

$$\left.\frac{\partial L(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} = \left.\frac{\partial \|\boldsymbol{b} - A\boldsymbol{x}\|^2}{\partial x_i}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} = 0$$

  - 위의 관계식으로부터 $\hat{\boldsymbol{x}}$가 만족해야 하는 등식 유도 가능

II.2 The Normal Equations $A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b}$

- 다변수함수 $f(x_1, x_2, \cdots, x_n)$의 gradient
  - 다변수함수 $f(x_1, x_2, \cdots, x_n)$의 값은 스칼라임
  - 다변수함수 f의 각 성분 변수에 대해 편미분을 한 결과를 성분으로 가지는 벡터를 주어진 함수의 gradient 라고 함

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \end{bmatrix} \Leftrightarrow \frac{\partial f}{\partial x_i} = (\nabla f)_i$$

  - 따라서 gradient 는 일종의 벡터임

- 벡터의 내적 형태로 표현된 스칼라식의 각 성분 변수들에 대한 미분
  - Notation: $(\boldsymbol{v})_i$는 벡터 $\boldsymbol{v}$의 i번째 성분, S 는 $A^T A$의 형태를 포함한 대칭 행렬
  - $\left(\nabla(\boldsymbol{x}^T \boldsymbol{x})\right)_i = \frac{\partial}{\partial x_i}(\boldsymbol{x}^T \boldsymbol{x}) = \frac{\partial}{\partial x_i}(x_1^2 + \cdots + x_n^2) = 2x_i = 2(\boldsymbol{x})_i$
    ➔ $\nabla(\boldsymbol{x}^T \boldsymbol{x}) = 2\boldsymbol{x}$
  - $\left(\nabla(\boldsymbol{x}^T S\boldsymbol{x})\right)_i = \frac{\partial}{\partial x_i}(\boldsymbol{x}^T S\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(\sum_j \sum_k S_{jk} x_j x_k\right) = 2\sum_k S_{ik} x_k = 2(S\boldsymbol{x})_i$
    ➔ $\nabla(\boldsymbol{x}^T S\boldsymbol{x}) = 2S\boldsymbol{x}$
  - $\boldsymbol{c}$는 상수값으로 이루어진 벡터이고, $\boldsymbol{c}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{c}$
  - $\left(\nabla(\boldsymbol{c}^T \boldsymbol{x})\right)_i = \frac{\partial}{\partial x_i}(\boldsymbol{c}^T \boldsymbol{x}) = \frac{\partial}{\partial x_i}(\boldsymbol{x}^T \boldsymbol{c}) = \frac{\partial}{\partial x_i}(c_1 x_1 + \cdots + c_n x_n) = c_i = (\boldsymbol{c})_i$
    ➔ $\nabla(\boldsymbol{c}^T \boldsymbol{x}) = 2\boldsymbol{c}$
  - 벡터 $\boldsymbol{x}^T = [x_1 \quad \cdots \quad x_n]$의 각각에 대한 미분은 다항식의 미분을 일반화한 형태임

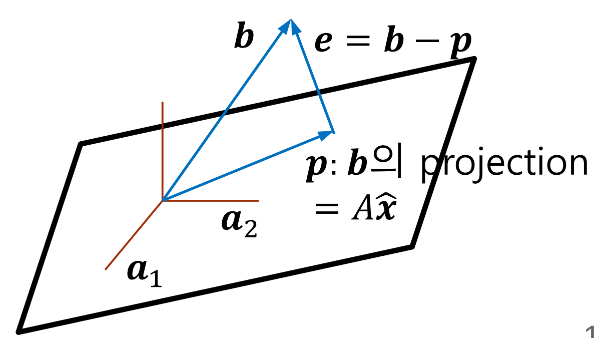
II.2 The Normal Equations $A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b}$

- $\|\boldsymbol{b} - A\boldsymbol{x}\|^2$의 최소화
  - $L(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \boldsymbol{x}^T A^T A\boldsymbol{x} - \boldsymbol{b}^T A\boldsymbol{x} - \boldsymbol{x}^T A^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b}$
  - 모든 x_1에 대해 $\frac{\partial \|\boldsymbol{b}-A\boldsymbol{x}\|^2}{\partial x_i} = 0$ 이 만족 $\Leftrightarrow \nabla\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \boldsymbol{0}$
  - $\frac{\partial}{\partial x_i}(\boldsymbol{x}^T A^T A\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(\sum_j \sum_k (A^T A)_{jk} x_j x_k\right) = 2\sum_k (A^T A)_{ik} x_k = 2(A^T A\boldsymbol{x})_i$
    $\Rightarrow \nabla(\boldsymbol{x}^T A^T A\boldsymbol{x}) = 2A^T A\boldsymbol{x}$
  - $\frac{\partial}{\partial x_i}(-\boldsymbol{b}^T A\boldsymbol{x} - \boldsymbol{x}^T A^T \boldsymbol{b}) = -2\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^T (A^T \boldsymbol{b})\right) = -2(A^T \boldsymbol{b})_i$
    $\Rightarrow \nabla(-2\boldsymbol{x}^T A^T \boldsymbol{b}) = -2A^T \boldsymbol{b}$
  - $\frac{\partial}{\partial x_i}(\boldsymbol{b}^T \boldsymbol{b}) = ?$
  - 모든 x_1에 대해 $\frac{\partial \|\boldsymbol{b}-A\boldsymbol{x}\|^2}{\partial x_i} = 0$ 이 만족 $\Leftrightarrow \nabla\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \boldsymbol{0}$

$$\frac{\partial \|\boldsymbol{b}-A\boldsymbol{x}\|^2}{\partial x_i} = 2(A^T A\boldsymbol{x})_i - 2(A^T \boldsymbol{b})_i = 0 \Rightarrow A^T A\boldsymbol{x} - A^T \boldsymbol{b} = \boldsymbol{0}$$

Ⅱ.2 $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$의 기하학적 의미

- 부분 공간에 projection하는 방법
  - 예를 들어 $\mathbf{R}^3$공간에서 벡터 $\boldsymbol{b}$를 어떤 평면에 가장 가까운 점에 projection하는 방법은?
  - 만약 벡터 $\boldsymbol{a}_1$와 $\boldsymbol{a}_2$ 가 평면의 기저이면, 평면은 이 두 벡터를 열로 가지는 행렬 $A = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2]$의 열벡터 공간에 해당한다.
  - 이 평면에 projection된 벡터 $\boldsymbol{p}$는 $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 = A\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$의 형태로 나타남.
  - x_1, x_2를 구하는 방법?
    - $\boldsymbol{b} - \boldsymbol{p}$를 이어주는 선과 평면의 모든 기저 벡터 $\boldsymbol{a}_i$가 직교한다는 조건을 이용. $\boldsymbol{b} - \boldsymbol{p} = \boldsymbol{b} - A\hat{\boldsymbol{x}}$
  - $\begin{cases} \boldsymbol{a}_1^T(\boldsymbol{b} - A\hat{\boldsymbol{x}}) = 0 \\ \boldsymbol{a}_2^T(\boldsymbol{b} - A\hat{\boldsymbol{x}}) = 0 \end{cases} \Rightarrow A^T(\boldsymbol{b} - A\hat{\boldsymbol{x}}) = \boldsymbol{0}$
  - 따라서 $\|\boldsymbol{b} - A\hat{\boldsymbol{x}}\|^2$를 최소로 만드는 $\hat{\boldsymbol{x}}$를 찾았다는 것은 벡터 $\boldsymbol{b}$의 행렬 A의 열벡터 부분 공간 성분을 $\boldsymbol{p}$라고 할 때, $A\hat{\boldsymbol{x}} = \boldsymbol{p}$를 만족하는 해를 찾은 것으로 해석 가능함



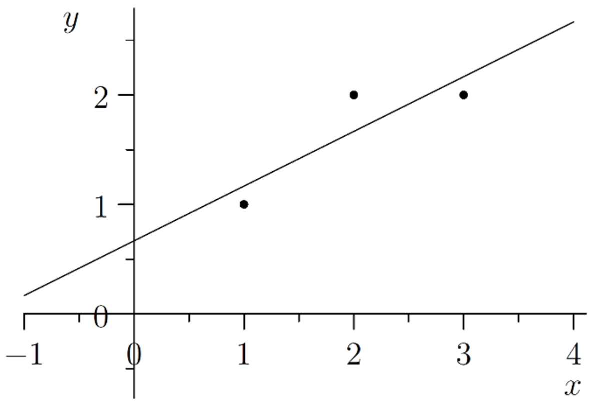
II.2 The Normal Equations $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$

- 최소 자승법 (least-squares)
  - 예를 들어 3개의 데이터$(t, b) = \{(1,1), (2,2), (3,2)\}$를 가지고 있다고 가정
  - 이 데이터에 fitting할 수 있는 직선 $b = C + Dt$를 찾고자 하면 다음과 같은 연립방정식이 가능함
  - $\begin{cases} C + D = 1 \\ C + 2D = 2 \\ C + 3D = 2 \end{cases} \Rightarrow A\boldsymbol{x} = \begin{bmatrix} \quad \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix} = \boldsymbol{b}$
  - 일반적으로 해가 존재하는 문제인가?
    - 아니오. 벡터 $\boldsymbol{b}$가 행렬 A의 열벡터 공간에 포함되지 않는 경우 해가 없음
  - 역행렬과 유사한 것을 만드는 방법은?
    - 행렬 A의 열들이 모두 독립이면, A^TA의 역행렬이 존재하고 $(A^TA)\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$을 이용하여 $\hat{\boldsymbol{x}} = (A^TA)^{-1}A^T\boldsymbol{b}$을 구할 수 있음

II.2 The Normal Equations $A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$

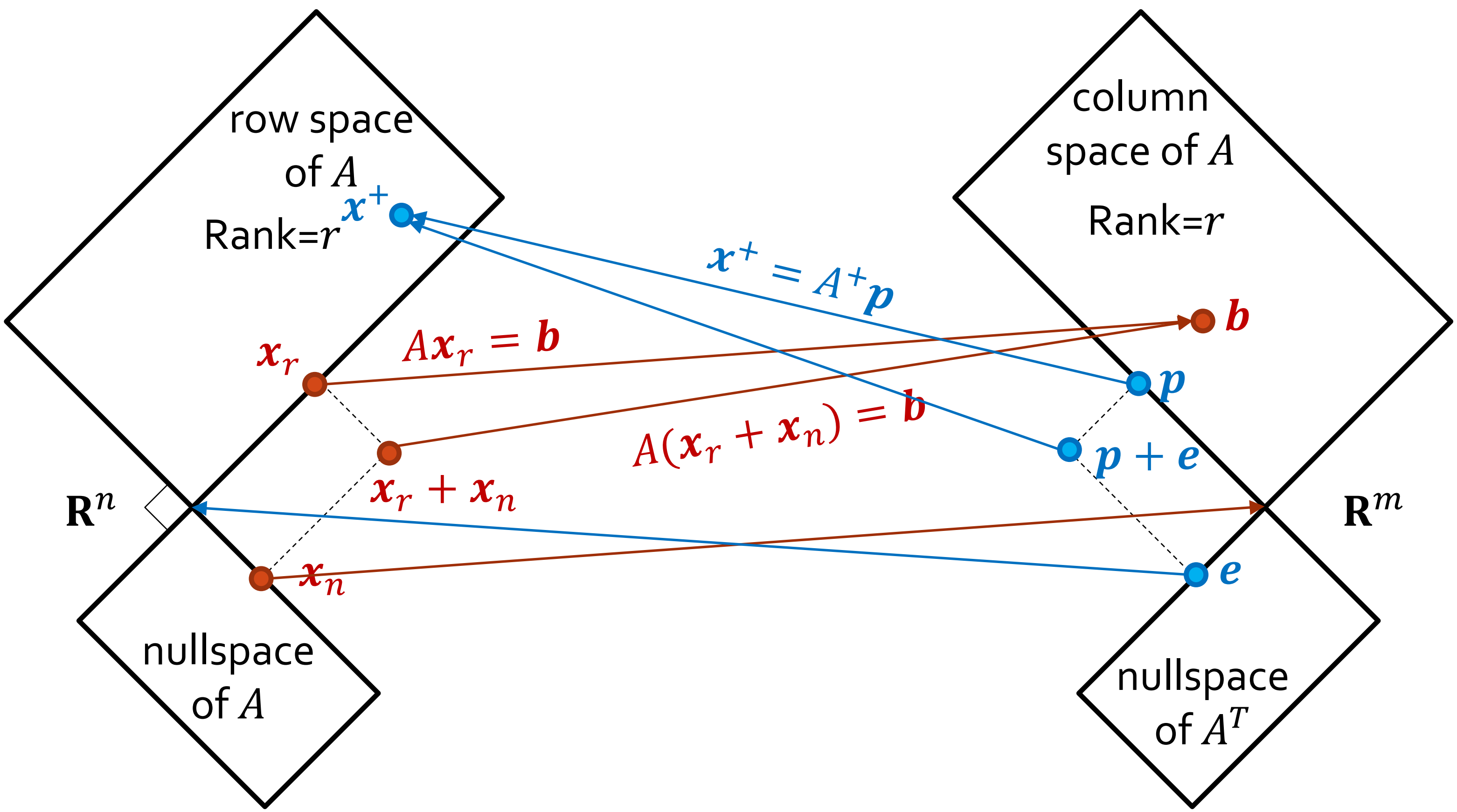
- 최소 자승법 (least-squares)
  - $A\boldsymbol{x} = \boldsymbol{b} \rightarrow A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$
  - $A^T A\hat{\boldsymbol{x}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}$

    $= A^T\boldsymbol{b} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}$
  - $(A^T A)^{-1} = \frac{1}{42-36} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix}$
  - $\hat{\boldsymbol{x}} = (A^T A)^{-1} A^T\boldsymbol{b} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 5 \\ 11 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/2 \end{bmatrix} = \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}$
  - $b = C + Dt = \frac{2}{3} + \frac{1}{2}t$



II.2 A^+ is the Pseudoinverse of A

- 행렬의 pseudoinverse A^+(의사역행렬, 유사역행렬)
  - 차원이 r인 열벡터 공간의 벡터를 행벡터 공간의 해당하는 벡터로 mapping해주는 행렬
  - 행벡터 공간의 벡터와 열벡터 공간의 벡터 간에는 일대일 대응 관계 존재



II.2 A^+ is the Pseudoinverse of A

- 유사역행렬
  - 행렬 A 의 SVD가 $A = U\Sigma V^T$일 때, 행렬 A의 유사역행렬은 $A^+ = V\Sigma^+ U^T$로 정의되고, Σ와 Σ^+는 아래와 같은 관계를 가짐
  - $\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}, \Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & & & \\ & \ddots & & & \\ & & 1/\sigma_r & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}$
  - 행렬 A 의 크기가 $m \times n$일 때 SVD를 계산하면 Σ의 크기 역시 $m \times n$이고, 역으로 유사역행렬 A 와 Σ^+는 $n \times m$ 이 됨.
  - 만약 행렬 A의 모든 열이 독립이면, $A^+ = (A^T A)^{-1} A^T$ ➔ $A^+ A = I$
    - 이 때 AA^+의 크기는 rank보다 크므로, 단위 행렬이 되지 않음
  - 만약 행렬 A의 모든 행이 독립이면, $A^+ = A^T (AA^T)^{-1}$ ➔ $AA^+ = I$
    - 이 때 A^+A의 크기는 rank보다 크므로, 단위 행렬이 되지 않음

II.2 A^+ is the Pseudoinverse of A

- $\boldsymbol{x}^+ = A^+\boldsymbol{b}$는 loss function $L(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$를 최소화시키는 해 중에서 norm이 최소인 해임
  - (증명) 임의의 행렬 A는 SVD에 의해 $A = U\Sigma V^T$의 형태로 분해가 가능하므로, 다음과 같이 표현 가능
  - $\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \|\boldsymbol{b} - U\Sigma V^T\boldsymbol{x}\|^2$
  - U는 orthogonal 행렬이므로 U 또는 $U^T = U^{-1}$를 벡터에 곱하더라도 길이가 보존됨
  - $\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \|\boldsymbol{b} - U\Sigma V^T\boldsymbol{x}\|^2 = \|U^T(\boldsymbol{b} - U\Sigma V^T\boldsymbol{x})\|^2 = \|U^T\boldsymbol{b} - \Sigma V^T\boldsymbol{x}\|^2$
  - $U^T\boldsymbol{b}$와 $V^T\boldsymbol{x}$는 둘다 벡터이므로 $\boldsymbol{c} = U^T\boldsymbol{b}$와 $\boldsymbol{w} = V^T\boldsymbol{x}$로 나타내면, 행렬의 rank가 2인 경우 다음과 같음

$$U^T\boldsymbol{b} - \Sigma V^T\boldsymbol{x} = \boldsymbol{c} - \Sigma\boldsymbol{w} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{bmatrix} - \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & 0 & \\ & & & \ddots \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{bmatrix} - \begin{bmatrix} \sigma_1 w_1 \\ \sigma_2 w_2 \\ 0 \\ \vdots \end{bmatrix}$$

  - 따라서 벡터 $\boldsymbol{w}$는 아래와 같을 때, $L(\boldsymbol{x})$을 최소로 만듦.

$$\boldsymbol{w}_0 = \begin{bmatrix} c_1/\sigma_1 \\ c_2/\sigma_2 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \\ & & 0 & \\ & & & \ddots \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{bmatrix} = \Sigma^+\boldsymbol{c}$$

  - V 는 orthogonal행렬이고 $\boldsymbol{w} = V^T\boldsymbol{x}$의 관계로부터 $\boldsymbol{w}_0$에 해당하는 $\boldsymbol{x}_0$는 아래와 같이 $\boldsymbol{x}^+$가 됨.

$$\boldsymbol{x}_0 = V\boldsymbol{w}_0 = V\Sigma^+\boldsymbol{c} = V\Sigma^+U^T\boldsymbol{b} = A^+\boldsymbol{b} = \boldsymbol{x}^+$$

  - A^+는 $\mathbf{R}^m$ 공간의 벡터를 행벡터 부분 공간에 속하는 벡터로 변환하므로 $\boldsymbol{x}^+$는 행벡터 공간에 속하고 nullspace 부분 공간에 속하는 임의의 벡터 $\boldsymbol{x}_n$과는 직교함.
  - $\|\boldsymbol{b} - A(\boldsymbol{x}^+ + \boldsymbol{x}_n)\|^2 = \|\boldsymbol{b} - A\boldsymbol{x}^+\|^2$ 이므로 $\boldsymbol{x}^+ + \boldsymbol{x}_n$ 역시 $L(\boldsymbol{x})$를 최소화시킴.
  - 하지만, 아래와 같은 관계로 인해 $\boldsymbol{x}^+$의 norm이 최소값을 가짐

$$\|\boldsymbol{x}^+ + \boldsymbol{x}_n\|^2 = \|\boldsymbol{x}^+\|^2 + \|\boldsymbol{x}_n\|^2 > \|\boldsymbol{x}^+\|^2$$

II.2 A^+ is the Pseudoinverse of A

- Example) 유사역행렬 구하기

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} = U\Sigma V^T = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \end{bmatrix}$$

$$A^+ = \mathrm{V}\Sigma^+ U^T = \begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & -1/\sqrt{2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/2 \\ 1/3 & -1/2 \\ 1/3 & 0 \end{bmatrix}$$

$$AA^+ = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1/3 & 1/2 \\ 1/3 & -1/2 \\ 1/3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A^+A = \begin{bmatrix} 1/3 & 1/2 \\ 1/3 & -1/2 \\ 1/3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 5/6 & -1/6 & 1/3 \\ -1/6 & 5/6 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \rightarrow \boldsymbol{v}_1 = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

$$(A^+A)\boldsymbol{v}_1 = \boldsymbol{v}_1,\ (A^+A)\boldsymbol{v}_2 = \boldsymbol{v}_2,\ (A^+A)\boldsymbol{v}_3 = \mathbf{0}$$

(선형대수) 9.2 Norms and Condition Number

- 행렬의 condition number
  - $A\boldsymbol{x} = \boldsymbol{b}$와 같은 등식에서 A의 성분의 값이 조금 바뀌었을 때, $\boldsymbol{x}$의 모든 성분의 값도 조금만 바뀔 경우 이 행렬은 well-conditioned matrix라 부르고, 작은 변화에도 크게 변할 경우 ill-conditioned matrix라고 불림
    - $\begin{bmatrix} 400 & -201 \\ -800 & 401 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ -200 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -100 \\ -200 \end{bmatrix}$
    - $\begin{bmatrix} 401 & -201 \\ -800 & 401 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ -200 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 40000 \\ 79800 \end{bmatrix}$
  - 행렬의 condition number $\mathrm{cond}[A]$는 다음과 같이 행렬의 norm을 이용해 정의되고, 이 값이 작은 경우에는 well-conditioned이고 큰 경우에는 ill-conditioned 가 됨

$$\mathrm{Cond}[A] = \|A\|\|A^{-1}\|$$

  - 행렬의 norm 중 아래의 spectrum norm (l^2-norm) 이 사용된 경우 $\|A\|$는 σ_1이고 $\|A^{-1}\|$는 $1/\sigma_r$이므로, 일반적으로 condition number는 대략 singular value의 최대값과 최소값의 비율에 해당함

$$\|A\|_2 = \max \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1$$

  - Ill-conditioned matrix의 경우 역행렬 계산이 필요한 경우 컴퓨터의 round-off와 같은 오차로 인해 계산 과정의 stability 가 많은 영향을 받게 되고 부정확도가 증가

    ➔ condition number 가 작은 행렬을 계산하는 것이 유리

II.2 The Third Way to Compute $\hat{x}$: Gram-Schmidt

- Gram-Schmidt 직교화를 통한 해의 계산
  - 행렬 A의 열들이 모두 독립인 경우에도, $\hat{\boldsymbol{x}} = (A^TA)^{-1}A^T\boldsymbol{b}$의 계산의 어려움이 존재
    - 열벡터들이 orthogonal 하지 않으므로 A^TA가 대각 행렬이 아님 ➔ 역행렬 계산 필요
    - A^TA의 condition number 는 A의 condition number 의 제곱 정도에 해당하여 행렬 A가 ill-conditioned matrix 의 경우에는 좋은 방법이 아님
  - Orthogonal 행렬
    - Orthogonal 행렬 Q의 모든 singular value는 1임. Why?
    - Orthogonal 행렬 Q의 condition number 는 1임
  - 주어진 행렬 A를 $A = QR$의 형태로 분해
    - $[\boldsymbol{a} \quad \boldsymbol{b} \quad \boldsymbol{c}] = [\boldsymbol{q}_1 \quad \boldsymbol{q}_2 \quad \boldsymbol{q}_3]\begin{bmatrix} \boldsymbol{q}_1^T\boldsymbol{a} & \boldsymbol{q}_1^T\boldsymbol{b} & \boldsymbol{q}_1^T\boldsymbol{c} \\ & \boldsymbol{q}_2^T\boldsymbol{b} & \boldsymbol{q}_2^T\boldsymbol{c} \\ & & \boldsymbol{q}_3^T\boldsymbol{c} \end{bmatrix}$
  - $A\hat{\boldsymbol{x}} = \boldsymbol{b}$ ➔ $R = Q^TA$이므로$Q^TA\hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$ ➔ $R\hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$
  - R 은 upper-right triangle의 형태이므로 $\hat{\boldsymbol{x}} = R^{-1}Q^T\boldsymbol{b}$를 쉽게 얻을 수 있음