4. $f:\mathbb{R}^d \longrightarrow \mathbb{R}^p$, $f(x) = W^T x$

$W \in \mathbb{R}^{d\times p}$ $\quad x \in \mathbb{R}^d \quad y \in \mathbb{R}^p$

$$\{(x_i, y_i)\}_{i=1}^{n}$$

Let $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times p}$

$$J(W) = \sum_{i=1}^{n} \| W^T x_i - y_i \|_2^2$$

Consider i^{th} instance

$$x_i = \langle x_{i1}, \dots, x_{id} \rangle^T \xrightarrow{\text{extend w/ bias}}$$

extend w/ bias $\langle 1, x_{i1}, \dots, x_{id} \rangle^T$

$$W = \begin{pmatrix} W_{00} & \cdots & W_{0p} \\ & \vdots & \\ W_{d0} & \cdots & W_{dp} \end{pmatrix}$$

$$y_i = \langle y_{i1}, \dots, y_{ip} \rangle^T$$

$$\Rightarrow W^T x_i - y_i = \langle z_{i1}, \dots, z_{ip} \rangle^T$$

$$\| W^T x_i - y_i \|_2^2 = \left(z_{i1}^2 + \dots + z_{ip}^2 \right)$$

We'll convert $J(W)$ in terms of X, Y:

$$y_i = w_{0i} + w_{1i} x_{i1} + \ldots + w_{di} x_{id} + \underbrace{\varepsilon_i}_{\text{error}}$$

Thus,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & & \vdots & \\ \vdots & & \vdots & \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ & & \vdots & \\ y_{np} & \cdots & & y_{np} \end{pmatrix}$$

Get L_2 norm equivalent:

Consider $(XW - Y)^T (XW - Y)$

Now $XW - Y \in \mathbb{R}^{n \times p}$

$$= \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ & & \vdots & \\ & & \vdots & \\ z_{n1} & & \cdots & z_{np} \end{pmatrix}$$

Where $z_{ij} = (w_{0j} + w_{1j} x_{i1} + \ldots + w_{dj} x_{id})$

Consider $(Xw-Y)(Xw-Y)^T$

The summation

$$\sum_{i=1}^{n} \| w^T x_i - y_i \|_2^2$$

occurs on the diagonal of

$$(Xw-y)(Xw-y)^T$$

$$\Rightarrow J(w) = tr\left[(Xw-y)(Xw-y)^T\right]$$

$$= tr(yy^T) - 2tr((Xw)^T y) + tr((Xw)^T Xw)$$

$$J(w) = tr(y^T y) - 2tr(y^T Xw) + tr(w^T X^T X w)$$

Closed form:

$$\frac{\partial J(w)}{\partial w} = -2\frac{\partial}{\partial w}(y^T Xw) + \frac{\partial}{\partial w}(w^T X^T X w)$$

$$0 = -2X^T y + 2(X^T X w)$$

$$X^T y = X^T X w$$

$$\hat{w} = (X^T X)^{-1} X^T y$$

4 b)

$$Y \in \mathbb{R}^{m \times p}, \quad W \in \mathbb{R}^{d \times p}, \quad X \in \mathbb{R}^{M \times d}.$$

Closed form

$$\hat{w} = (X^T X)^{-1} X^T y$$

Consider i^{th} column of Y

$$y_i = \begin{pmatrix} y_{1i} \\ \vdots \\ y_{Mi} \end{pmatrix}$$

Now, $(X^T X)^{-1} X^T = H \in \mathbb{R}^{d \times M}$
and this matrix H is similar for a linear regression problem and a multivariate linear regression problem

For $Hy_i \in \mathbb{R}^{d \times 1}$, which corresponds to the i^{th} column of $\hat{w}$.

Thus, for y_i, $i = 1, \dots, p$, we get the p columns of $\hat{w}$

$\hat{w}_i = i^{th}$ column of $\hat{w}$

$$= \begin{pmatrix} w_{1i} \\ \vdots \\ w_{di} \end{pmatrix} = \begin{pmatrix} (h_{11} y_{1i} + \dots + h_{1M} y_{Mi}) \\ \vdots \\ (h_{d1} y_{1i} + \dots + h_{dM} y_{Mi}) \end{pmatrix}$$

$$\Rightarrow \quad \hat{W} = (X^T X)^{-1} X^T Y$$

$$\Longleftrightarrow \quad \hat{w}_i = (X^T X)^{-1} X^T y_i \qquad i = 1, \ldots, p$$

where $\hat{w}_i$, y_i are i^{th} column of $\hat{W}$, Y.

$\Rightarrow$ The multivariate Regression can be broken down into p separate linear regression problems.

This solution isn't appropriate if there exists any sort of relationship b/t the dependant (y) variables.

ex. Using time, season, weather, temperature predict the number of people visiting a bike shop for the day AND the revenue for the day.

In this example, revenue (y_2)s may depend on the number of people predicted to visit the shop (y_1), and thus, treating y_1 and y_2 as independent and solving using

$$\hat{W} = (X^T X)^{-1} X^T Y$$

may not be appropriate.

4c) $W \in \mathbb{R}^{d \times p}$

Let W be of low rank R.
Since every finite-dimensional matrix has a rank decomposition

$$W = AB, \quad A \in \mathbb{R}^{d \times R} \quad B \in \mathbb{R}^{R \times p}$$

Thus, low rank regression:

$$\min_{A,B} \| Y - X^T AB \|^2, \quad X \in \mathbb{R}^{d \times m}, \quad Y \in \mathbb{R}^{m \times p}$$

$$\equiv \min_{A,B} \| Y - (A^T X)^T B \|^2$$

Thus, low-rank regression can be viewed as standard linear regression but in the $\overline{X} = A^T X \in \mathbb{R}^{R \times m}$ subspace.

Consider i^{th} column of Y

$$= \forall_{i=1,\dots,p} \min_{A,B} \| y_i - (A^T x_i)^T B \|$$

$A^T x_i \in \mathbb{R}^{R \times 1}$ which is a R-dimensional subspace let $\overline{X}_i = A^T x_i$

Thus, $y_i^T = \overline{X}_i^T B + \varepsilon_i$

Since this is akin to linear regression in a R-dimensional subspace of $\mathbb{R}^p$, this leaves a $(p-R)$-dimensional of just error terms

$$\varepsilon_{i=(p-R),\ldots,p} \sim \mathcal{N}(0, \sigma^2 I)$$

$\Rightarrow \exists\ (p-R)$-dimensional subspace of $\mathbb{R}^p$ containing only noise.