MFDNN

Homework 1

24/03/08

Using underline for column vectors
Using CAPITAL with no subscripts for matrices
All others are scalars.

$\underline{\theta} \in \mathbb{R}^p$

1. (a) $$\frac{\partial}{\partial \theta_j} \ell_i(\underline{\theta}) = \frac{\partial}{\partial \theta_j}\left(\frac{1}{2}(\underline{x_i}^T\underline{\theta} - y_i)^2\right) = \frac{\partial}{\partial \theta_j}\left(\frac{1}{2}\left(\left(\sum_{n=1}^{p} x_{in}\theta_n\right) - y_i\right)^2\right)$$

$$= x_{ij}(\underline{x_i}^T\underline{\theta} - y_i)$$

$$\underline{\nabla}_\theta \ell_i(\underline{\theta}) = \sum_{j=1}^{p}\left(\frac{\partial}{\partial \theta_j}\ell_i(\underline{\theta})\right)\underline{e}_j$$

$$= \sum_{j=1}^{p}\left(x_{ij}(\underline{x_i}^T\underline{\theta} - y_i)\right)\underline{e}_j$$

$$= (\underline{x_i}^T\underline{\theta} - y_i)\underline{x_i} \quad \blacksquare$$ since $\sum_{j=1}^{p} x_{ij}\underline{e}_j = \begin{pmatrix} x_{i1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_{i2} \\ \vdots \\ 0 \end{pmatrix} + \dots + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_{ip} \end{pmatrix} = \underline{x_i}$

$\in \mathbb{R}^p$

(b) $$\frac{\partial}{\partial \theta_j}\mathcal{L}(\underline{\theta}) = \frac{\partial}{\partial \theta_j}\left(\frac{1}{2}(X\underline{\theta} - \underline{y})^T(X\underline{\theta} - \underline{y})\right) = \frac{\partial}{\partial \theta_j}\left(\frac{1}{2}\left(\sum_{i=1}^{p}(\underline{x_{:,i}}\theta_i) - \underline{y}\right)^2\right)$$ using given

$$= \underline{x_{:,j}}^T\left(\sum_{i=1}^{p}(\underline{x_{:,i}}\theta_i) - \underline{y}\right)$$

$\underline{y}, \underline{x_{:,i}} \in \mathbb{R}^{\text{“}N\times 1\text{”}}$

$$= \underline{x_{:,j}}^T(X\underline{\theta} - \underline{y})$$ using given again

$\underline{x_{:,i}}^T \in \mathbb{R}^{\text{“}1\times N\text{”}}$

$$\underline{\nabla}_\theta \mathcal{L}(\underline{\theta}) = \sum_{j=1}^{p}\underline{e}_j\left(\overbrace{\underline{x_{:,j}}^T(X\underline{\theta} - \underline{y})}^{\text{scalar}}\right)$$

$\underline{e}_j \in \mathbb{R}^{p\times 1}$

$$= \sum_{j=1}^{p}\left[(\underline{e}_j\underline{x_{:,j}}^T)\overbrace{(X\underline{\theta} - \underline{y})}^{\text{“}N\times 1\text{”}}\right]$$ since matrix multiplication is associative

$$= \left(\sum_{j=1}^{p}\underline{e}_j\underline{x_{:,j}}^T\right)(X\underline{\theta} - \underline{y})$$

$$= X^T(X\underline{\theta} - \underline{y}) \quad \blacksquare$$

$\in \mathbb{R}^p$

since $\underline{x_{:,j}}^T$ is jth column of X as a row vector
i.e. $(x_{1j}, x_{2j}, \dots, x_{Nj})$

$$\underline{e}_1\underline{x_{:,1}}^T = \overset{\text{“}p\times 1\text{”}}{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}\overset{1\times N}{(x_{11}, x_{21}, \dots, x_{N1})} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{p\times N}$$

$$\Rightarrow \sum_{j=1}^{p}\underline{e}_j\underline{x_{:,j}}^T = X^T$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \ddots & & \vdots \\ x_{N1} & \cdots & & x_{Np} \end{pmatrix} \in \mathbb{R}^{N\times p}$$

2. $\theta^{k+1} = \theta^k - \alpha f'(\theta^k)$ $\qquad f(x) = x^2/2$

$\theta^{k+1} = \theta^k - \alpha\theta^k$ $\qquad f'(x) = x$

$\theta^{k+1} = \theta^k(1-\alpha)$

$\frac{\theta^{k+1}}{\theta^k} = 1-\alpha$ valid for all k since $\theta^0 \neq 0$

$\lim_{k\to\infty}\left|\frac{\theta^{k+1}}{\theta^k}\right| = \lim_{k\to\infty}|1-\alpha|$ By the ratio test, the sequence θ^k diverges (i.e. $|\theta^k|\to\infty$) if $\lim_{k\to\infty}|1-\alpha| > 1$.

$\Rightarrow |1-\alpha| > 1 \Leftrightarrow 1-\alpha > 1$ or $1-\alpha < -1$

$\Leftrightarrow 0 > \alpha$ or $2 < \alpha$

Since α is positive, $\alpha > 2$ only relevant condition

i.e. θ sequence diverges if $\alpha > 2$ ∎

3. $f(\underline{\theta}) = \frac{1}{2}\|X\underline{\theta} - \underline{Y}\|^2$

$\nabla_\theta f(\underline{\theta}) = X^T(X\underline{\theta} - \underline{Y})$ from problem 1.

$\underline{0} = X^T(X\underline{\theta} - \underline{Y})$

$\underline{0} = X^TX\underline{\theta} - X^T\underline{Y} \Rightarrow \underline{\theta}^* = (X^TX)^{-1}X^T\underline{Y}$ is optimal solution (assuming X^TX invertible)

NB: Also implies $X^TX\underline{\theta}^* = X^T\underline{Y}$

Consider, $\underline{\theta}^{k+1} - \underline{\theta}^* = \underline{\theta}^k - \alpha\nabla f(\underline{\theta}^k) - \underline{\theta}^*$

"error vector"

$$
\begin{aligned}
&= \underline{\theta}^k - \alpha\left(X^T(X\underline{\theta}^k - \underline{Y})\right) - \underline{\theta}^* \\
&= \underline{\theta}^k - \alpha X^TX\underline{\theta}^k + \alpha X^T\underline{Y} - \underline{\theta}^* \\
&= \underline{\theta}^k - \alpha X^TX\underline{\theta}^k + \alpha X^TX\underline{\theta}^* - \underline{\theta}^* \\
&= \underline{\theta}^k - \underline{\theta}^* - \alpha X^TX(\underline{\theta}^k - \underline{\theta}^*) \\
&= (I - \alpha X^TX)(\underline{\theta}^k - \underline{\theta}^*)
\end{aligned}
$$

Set $A := I - \alpha X^TX$

$\Rightarrow$ The error vector will grow/diverge if $\rho(A) > 1$

(aside: to see this, consider $\underline{x} = \underline{\theta}^k - \underline{\theta}^*$ and $A = I - \alpha X^TX$.

Let A have unique eigenvectors $\underline{v}_1, \dots, \underline{v}_n$ alongside eigenvals $\lambda_1, \dots, \lambda_n$ $(n \leq p)$

$\underline{x}$ can be expressed in basis of eigenvectors: $\underline{x} = c_1\underline{v}_1 + \dots + c_n\underline{v}_n$

In this case, $A\underline{x} = A(c_1\underline{v}_1 + \dots + c_n\underline{v}_n)$

$$
\begin{aligned}
&= c_1 A\underline{v}_1 + \dots + c_n A\underline{v}_n \\
&= c_1\lambda_1\underline{v}_1 + \dots + c_n\lambda_n\underline{v}_n
\end{aligned}
$$

If we apply A again, i.e. $AA\underline{x} = \dots = c_1\lambda_1^2\underline{v}_1 + \dots + c_n\lambda_n^2\underline{v}_n$, it's clear that if $|\lambda_i| < 1$ then the components of $\underline{x}$, and hence $\|\underline{x}\|$ will shrink and, conversely, if $|\lambda_i| > 1$, $\|\underline{x}\|$ will grow/diverge. Hence we want to bound the largest eigenvalue (spectral radius) of A.)

Eigenvalues of A are $1 - \alpha\lambda_i$ where λ_i are eigenvalues of X^TX.

$\Rightarrow$ For divergence, $\exists i$ s.t. $|1 - \alpha\lambda_i| > 1$

In worst case, only one such i: $|1 - \alpha\rho(X^TX)| > 1$

$\rho(X^TX) := \max_i \{|\lambda_1|, \dots, |\lambda_i|\}$

$\Leftrightarrow 1 - \alpha\rho(X^TX) > 1$ or $1 - \alpha\rho(X^TX) < -1$

$\Leftrightarrow 0 > \alpha\rho(X^TX)$ or $2 < \alpha\rho(X^TX)$

N/A $\qquad 2/\rho(X^TX) < \alpha$

$\Rightarrow$ Error vector, by extension GD, diverges if $\alpha > 2/\rho(X^TX)$

(for most $\underline{\theta}^0$. There may be some $\underline{\theta}^0$ for instance with no component in the direction of a diverging eigenvector/value but but these are very specific in the context of $\mathbb{R}^n$ space)

$n-r+1 > 1$

$n > r$

5. (Workings) $\quad A \in \mathbb{R}^{(n-r+1)\times n} \quad \underline{x} \in \mathbb{R}^{n\times 1}$

$$\underline{Ax} = \underbrace{\begin{pmatrix} k_1 & \cdots & k_r & 0 & \cdots & & 0 \\ 0 & k_1 & \cdots & k_r & 0 & \cdots & 0 \\ 0 & 0 & k_1 & \cdots & k_r & 0 \cdots & 0 \\ \vdots & & \ddots & & & \ddots & \vdots \\ 0 & & \cdots & 0 & k_1 & \cdots & k_r \; 0 \\ 0 & & \cdots & 0 & 0 & k_1 \cdots & k_r \end{pmatrix}}_{n \text{ columns}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} k_1x_1 + k_2x_2 + \dots + k_rx_r \\ k_1x_2 + k_2x_3 + \dots + k_rx_{r+1} \\ \vdots \\ k_1x_{n-r} + k_2x_{n-r-1} + \dots + k_rx_{n-1} \\ k_1x_{n-r+1} + k_2x_{n-r+2} + \dots + k_rx_n \end{pmatrix} \in \mathbb{R}^{n-r+1}$$

$\downarrow$ in python indexes

$$k_0x_0 + \dots + k_{r-1}x_{r-1}$$
$$\vdots$$
$$k_0x_{n-r} + \dots + k_{r-1}x_{n-1}$$

$$\Rightarrow \underline{Ax}_i = \sum_{j=1}^{r} k_j \, x_{i+j-1}$$

$$\begin{pmatrix} k_0 \\ k_1 \\ \vdots \\ k_{r-1} \end{pmatrix} \underset{\text{element-wise multiplication}}{*} \begin{pmatrix} x_i \\ x_{i+1} \\ \vdots \\ x_{i+r-1} \end{pmatrix}$$

$x[i:i+r]$

i.e. dot product np.dot

$$\underline{A^T v} = \begin{pmatrix} k_1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & k_1 & 0 & & \vdots & \vdots \\ & \vdots & k_1 & & & \\ k_r & \vdots & & \ddots & 0 & 0 \\ 0 & k_r & \vdots & & k_1 & 0 \\ 0 & 0 & k_r & \ddots & & k_1 \\ \vdots & \vdots & \vdots & & k_r & \vdots \\ 0 & 0 & 0 & \cdots & 0 & k_r \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \\ v_{n-r+1} \end{pmatrix} = \begin{pmatrix} k_1v_1 \\ k_2v_1 + k_1v_2 \\ \vdots \\ k_rv_1 + k_{r-1}v_2 + \dots + k_1v_r \\ k_rv_2 + k_{r-1}v_3 + \dots + k_1v_{r+1} \\ \vdots \\ k_rv_{n-2r-1} + k_{r-1}v_{n-2r} + \dots + k_1v_{n-r} \\ k_rv_{n-2r} + k_{r-1}v_{n-2r+1} + \dots + k_2v_{n-r+1} \\ \vdots \\ k_rv_{n-r} + k_{r-1}v_{n-r+1} \\ k_rv_{n-r+1} \end{pmatrix}$$

$$\Uparrow \mathbb{R}^{n\times(n-r+1)} \qquad \Uparrow \mathbb{R}^{(n-r+1)\times 1} \qquad \Uparrow \mathbb{R}^{n}$$

$k \in \mathbb{R}^r$

reversed

$$\begin{pmatrix} k_r \\ k_{r-1} \\ \vdots \\ k_2 \\ k_1 \end{pmatrix} \cdot \begin{pmatrix} \left.\begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix}\right\} r-1 \\ v_1 \end{pmatrix} \begin{matrix} \cdots [i \\ \\ \\ \\ \cdots i+r-1] \end{matrix}$$

$\downarrow$ sliding

$$\begin{matrix} v_2 \\ \vdots \\ v_{n-r+1} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} r-1 \end{matrix}$$