2.

$$J(U,V) = \|UV - Y\|_F^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$$

Optimization

5. a. HW2

$\nabla J - ?$ $\quad U, V \in \mathbb{R}^{n\times n}$

$$\left\{ d\|X\|_F^2 = \langle 2X, dX\rangle \right.$$

① $d\|X\|_F^2 = d\left(\left(\langle X, X\rangle^{1/2}\right)^2\right) = d(\langle X, X\rangle) = \langle dX, X\rangle + \langle X, dX\rangle = \langle 2X, dX\rangle \Rightarrow$

$$dJ = \langle 2(UV-Y), d(UV-Y)\rangle + \frac{\lambda}{2}\left(\langle 2U, dU\rangle + \langle 2V, dV\rangle\right) =$$

$$= \langle 2(UV-Y), (dU\cdot V + U\cdot dV)\rangle + \lambda\langle U, dU\rangle + \lambda\langle V, dV\rangle =$$

$$= \langle 2(UV-Y), dU\cdot V\rangle + \langle 2(UV-Y), U dV\rangle + \lambda\langle U, dU\rangle + \lambda\langle V, dV\rangle =$$

② $\langle 2(UV-Y), dU\cdot V\rangle = \langle 2(UV-Y)V^T, dU\rangle$

$\langle X, Y\rangle := tr(X^T Y)$

$\langle X, Y\cdot A\rangle := tr(X^T Y A) \underset{tr AB = tr BA}{=} tr(A X^T Y) \underset{A^T B^T = (BA)^T}{=} tr((XA^T)^T Y) =: \langle XA^T, Y\rangle$

$$\langle X, YA\rangle = \langle XA^T, Y\rangle$$

③ $\langle 2(UV-Y), U dV\rangle = \langle 2U^T(UV-Y), dV\rangle$

$\langle X, AY\rangle := tr(X^T A Y) \underset{B^T A^T = (AB)^T}{=} tr((A^T X)^T, Y) =: \langle A^T X, Y\rangle$

$$\langle X, AY\rangle = \langle A^T X, Y\rangle$$

$$= \langle 2(UV-Y)V^T, dU\rangle + \langle 2U^T(UV-Y), dV\rangle + \lambda\langle U, dU\rangle + \lambda\langle V, dV\rangle$$

$$= \langle 2(UV-Y)\cdot V^T + \lambda U, dU\rangle + \langle 2U^T(UV-Y) + \lambda V, dV\rangle$$

$$\Rightarrow \quad \boxed{\nabla J = \begin{pmatrix} 2(UV-Y)V^T + \lambda U \\ 2U^T(UV-Y) + \lambda V \end{pmatrix}}$$

Briefly: $dJ = \langle 2(UV-Y), d(UV-Y)\rangle + \frac{\lambda}{2}(\langle 2U, dU\rangle + \langle 2V, dV\rangle) =$

$= \langle 2(UV-Y), dU\,V\rangle + \langle 2(UV-Y), U dV\rangle + \lambda\langle U, dU\rangle + \lambda\langle V, dV\rangle =$

$= \langle 2(UV-Y)V^T, dU\rangle + \langle 2U^T(UV-Y), dV\rangle + \lambda\langle U, dU\rangle + \lambda\langle V, dV\rangle =$

$= \langle 2(UV-Y)V^T + \lambda U, dU\rangle + \langle 2U^T(UV-Y) + \lambda V, dV\rangle$

$$f(w) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i w^T x_i}\right) \quad x_i \in \mathbb{R}^n, \; y_i \in \mathbb{R}$$

$$\boxed{\sigma = \frac{1}{1 + e^{-y_i w^T x_i}}} \Rightarrow e^{-y_i w^T x_i} = \frac{1}{\sigma} - 1 = \frac{1-\sigma}{\sigma}$$

$$\sigma'_w = (-1)(-y_i x_i) \cdot \frac{e^{-y_i w^T x_i}}{\left(1 + e^{-y_i w^T x_i}\right)^2} = y_i x_i \cdot \frac{1-\sigma}{\sigma} \cdot \frac{1}{\left(1 + \frac{1}{\sigma} - 1\right)^2} \Rightarrow$$

$$\boxed{\sigma'_w = y_i x_i (1-\sigma)\sigma}$$

$$f(w) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i w^T x_i}\right) = \sum_{i=1}^{m} \log\left(1 + \frac{1}{\sigma} - 1\right) = -\sum_{i=1}^{m} \log \sigma$$

$$\boxed{f(w) = -\sum_{i=1}^{m} \log \sigma}$$

$$\nabla f = -\sum_{i=1}^{m} \sigma' \cdot \frac{1}{\sigma} = -\sum_{i=1}^{m} y_i x_i (1-\sigma)\cancel{\sigma} \cdot \frac{1}{\cancel{\sigma}}$$

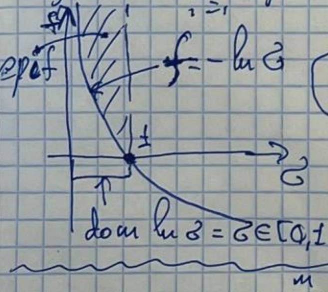
$$\boxed{\nabla f(w) = -\sum_{i=1}^{m} y_i x_i (1-\sigma)}$$

$$\nabla^2 f = -\sum_{i=1}^{m} y_i x_i (-\sigma') = \sum_{i=1}^{m} y_i^2 x_i x_i^T (1-\sigma)\sigma$$

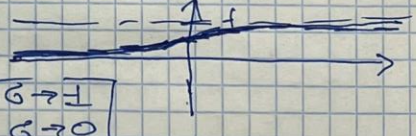
$$\boxed{f(w) = -\sum_{i=1}^{m} \log \sigma \; ; \; \nabla f(w) = -\sum_{i=1}^{m} y_i x_i (1-\sigma) ; \; \nabla^2 f(w) = \sum_{i=1}^{m} y_i^2 x_i x_i^T (1-\sigma)\cdot\sigma \quad \text{where } \sigma = \frac{1}{1 + e^{-y_i w^T x_i}}}$$

About convexity: If ∇f is monotonic: $\langle x-y, \nabla f(x) - \nabla f(y) \rangle \geq 0$ and dom f is convex, then f – is convex.

$$f(w) = -\sum_{i=1}^{m} \log \sigma, \quad \sigma = \sigma(-y_i w^T x_i) = \frac{1}{1 + e^{-y_i w^T x_i}} \quad x_i \in \mathbb{R}^n \; y_i \in \mathbb{R}$$



$\boxed{\sigma \in [0, 1]}$ in 2-d →



$$\begin{cases} -y_i w^T x_i \to +\infty & \sigma \to 1 \\ -y_i w^T x_i \to -\infty & \sigma \to 0 \end{cases}$$

dom $\ln \sigma = \sigma \in [0, 1]$

$$\langle \sigma(w) - \sigma(w_0), \nabla f(\sigma(w)) - \nabla f(\sigma(w_0)) \rangle =$$

$$= \langle \sigma(w) - \sigma(w_0), -\sum_{i=1}^{m} y_i x_i (1 - \sigma(w)) + \sum_{i=1}^{m} y_i x_i (1 - \sigma(w_0)) \rangle$$

$$= \langle \sigma(w) - \sigma(w_0), \sum_{i=1}^{m} (\sigma(w) - \sigma(w_0)) \rangle$$

σ – is monotonically increasing $\Rightarrow$ if $w \geq w_0 \Rightarrow \sigma(w) \geq \sigma(w_0) \Rightarrow \sigma(w) - \sigma(w_0) \geq 0$

$\langle +, + \rangle \Rightarrow \langle +, + \rangle \geq 0$

if $w < w_0 \Rightarrow \sigma(w) < \sigma(w_0) \Rightarrow \sigma(w) - \sigma(w_0) < 0$

$\langle -, - \rangle \Rightarrow \langle -, - \rangle \geq 0$

(I) dom f is a segment $\sigma \in [0, 1]$ is a convex set.

(II) $\langle \sigma(w) - \sigma(w_0), \nabla f(\sigma(w)) - \nabla f(\sigma(w_0)) \rangle \geq 0$

$\underline{\text{I}} + \underline{\text{II}} \Rightarrow$ f is a convex function

HW 2.5.b
Page 2
Optimization

Another way (way 2) to prove convexity:

$$H = \nabla^2 f(w) = \sum_{i=1}^{m} y_i^2 x_i x_i^T \left(1 - \sigma(-y_i w^T x_i)\right) \sigma(-y_i w^T x_i)$$

$$x_i \in \mathbb{R}^n,\ y_i \in \mathbb{R}$$

Let $y_i x_i = z_i$, $y_i x_i^T = z_i^T$

$$H = \sum_{i=1}^{m} \underbrace{z_i z_i^T}_{\substack{Z Z^T \\ \text{matrices}}} \underbrace{\left(1 - \sigma(-w^T z_i)\right) \sigma(-w^T z_i)}_{\text{diagonal matrix } D} \Rightarrow$$

where $d_{ii} = \left(1 - \sigma(-w^T z_i)\right) \sigma(-w^T z_i)$

$$\Rightarrow H = Z Z^T D = Z D Z^T$$

H is PSD (positive semidefinite) if for $\forall p \in \mathbb{R}^n$ $p^T H p \geq 0$.

Then
$$p^T H p = p^T Z D Z^T p = p^T Z D^{1/2} D^{1/2} Z^T p =$$
$$= (D^{1/2} Z^T p)^T (D^{1/2} Z^T p) = \langle D^{1/2} Z^T p, D^{1/2} Z^T p \rangle =$$
$$= \underbrace{\| D^{1/2} Z^T p \|}_{\geq 0} \text{ - Frobenius norm} \Rightarrow$$

$\Rightarrow$ our $H = \nabla^2 f(w)$ is PSD $\Rightarrow$

$\Rightarrow f(w)$ is convex

Compute the Jacobi matrix of the following function $f: \mathbb{R}^n \to \mathbb{R}^n$, $f(w)_j = \dfrac{e^{w_j}}{\sum_{k=1}^{n} e^{w_k}}$

5.c.
Page 1

This is softmax function $\mathbb{R}^n \to \mathbb{R}^n$

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} & \cdots & \frac{\partial f_1}{\partial w_n} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} & \cdots & \frac{\partial f_2}{\partial w_n} \\ \vdots & & \cdots & \\ \frac{\partial f_n}{\partial w_1} & \frac{\partial f_n}{\partial w_2} & \cdots & \frac{\partial f_n}{\partial w_n} \end{pmatrix}$$

Logarithmic derivative:

$$\frac{\partial \log(f_i)}{\partial w_j} = \frac{1}{f_i} \cdot \frac{\partial f_i}{\partial w_j} \Rightarrow$$

$$\frac{\partial f_i}{\partial w_j} = f_i \cdot \frac{\partial}{\partial w_j} \log(f_i) \quad [\mathrm{I}]$$

$$\log f_i = \log\left(\frac{e^{w_i}}{\sum_{k=1}^{n} e^{w_k}}\right) = w_i - \log\left(\sum_{k=1}^{n} e^{w_k}\right)$$

$$\frac{\partial}{\partial w_k} \log f_i = \underbrace{\frac{\partial w_i}{\partial w_k}} - \frac{\partial}{\partial w_k} \log\left(\sum_{k=1}^{n} e^{w_k}\right)$$

where $\dfrac{\partial w_i}{\partial w_k} = \begin{cases} 1 & \text{if } i = k \\ 0 & i \neq k \end{cases}$

$$\frac{\partial}{\partial w_k} \log f_i = \mathbb{1}\{i=j\} - \frac{1}{\sum_{k=1}^{n} e^{w_k}} \cdot \underbrace{\left(\frac{\partial}{\partial w_k} \sum_{l=1}^{n} e^{w_l}\right)} \quad (\mathrm{II})$$

$$\frac{\partial}{\partial w_k} \sum_{l=1}^{n} e^{w_l} = \frac{\partial}{\partial w_k}\left(e^{w_1} + e^{w_2} + \ldots + e^{w_k} + \ldots + e^{w_n}\right) =$$

$$= \frac{\partial}{\partial w_k}\left(e^{w_k}\right) = e^{w_k}$$

$$[\mathrm{I}]\ \frac{\partial f_i}{\partial w_j} = f_i \underbrace{\frac{\partial}{\partial w_j} \log(f_i)}_{\to [\mathrm{II}]} = f_i \cdot [1(i=j) - f_j]$$

$$\Rightarrow \boxed{\frac{\partial f_i}{\partial w_j} = f_i \cdot [\, 1\{i=j\} - f_j \,]}$$

For example for case $n=3$

$$J_3 = \begin{pmatrix} f_1(1-f_1) & -f_1 \cdot f_2 & -f_1 \cdot f_3 \\ -f_2 \cdot f_1 & f_2(1-f_2) & -f_2 \cdot f_3 \\ -f_3 \cdot f_1 & -f_3 \cdot f_2 & f_3 \cdot (1-f_3) \end{pmatrix}$$