MSBD 5004 HOMEWORK 4.

Name: Zheng Mingen
Stu no: 21126390

1. To solve $\nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \vec{0}$, we just need to solve $\begin{cases} \frac{\partial f}{\partial x_1} = 0 \\ \frac{\partial f}{\partial x_2} = 0 \end{cases}$

then all the solutions (x_1, x_2) is what we want to find.

(a) $$\begin{cases} \frac{\partial f}{\partial x_1} = 2(4x_1^2 - x_2)\cdot(8x_1) = 0 & ① \\ \frac{\partial f}{\partial x_2} = 2(4x_1^2 - x_2)\cdot(-1) = 0 & ② \end{cases}$$

① $\Longleftrightarrow$ $x_2 = 4x_1^2$ or $x_1 = 0$

② $\Longleftrightarrow$ $x_2 = 4x_1^2$,

for $\forall x_1 \in R$ (which certainly contains $x_1 = 0$)

when $x_2 = 4x_1^2$ ✓, ① is always true. i.e.

those vectors in the form of $(a, 4a^2)$ for $\forall a \in R$ are all critical points.

(b) $$\begin{cases} \frac{\partial f}{\partial x_1} = 6x_1x_2 = 0 & ① \\ \frac{\partial f}{\partial x_2} = 6x_2^2 - 12x_2 + 3x_1^2 = 0 & ② \end{cases}$$

① $\Longleftrightarrow$ $x_1 = 0$ or $x_2 = 0$

✓ when $x_1 = 0$, ② can be reformulated as $x_2^2 = 2x_2$.

solve it then we have $x_2 = 0$ or $x_2 = 2$.

i.e. $(0,0)$, $(0,2)$ are critical points.

✓ when $x_2 = 0$, ② can be reformulated as $3x_1^2 = 0$

Solve it then we have $x_1 = 0$, i.e. $(0,0)$ is a critical point.

Altogether, critical points are $(0,0)$ and $(0,2)$

(c) $$\begin{cases} \frac{\partial f}{\partial x_1} = 4(x_1 - 2x_2)^3 + 64x_2 = 0 & ① \\ \frac{\partial f}{\partial x_2} = 4(x_1 - 2x_2)^3\cdot(-2) + 64x_1 = 0 & ② \end{cases}$$

(in next page)

① $\Leftrightarrow$ $16x_2 = -(x_1 - 2x_2)^3$

② $\Leftrightarrow$ $-8x_1 = -(x_1 - 2x_2)^3$

thus $16x_2 = -8x_1$ i.e. $x_1 = -2x_2$

Substitute x_1 in ①, then we have:

$$4(-4x_2)^3 + 64x_2 = 0$$

$$64x_2 = 4 \cdot 64x_2^3$$

$$\Rightarrow x_2 = 0 \quad \text{or} \quad x_2 = \frac{1}{2} \quad \text{or} \quad x_2 = -\frac{1}{2}$$

$x_1 = -2x_2$

$$\Rightarrow x_1 = 0 \quad \text{or} \quad x_1 = -1 \quad \text{or} \quad x_1 = 1$$

✓

i.e $(0, 0)$, $(-1, \frac{1}{2})$, $(1, -\frac{1}{2})$ are all the critical points.

(d)
$$\begin{cases} \frac{\partial f}{\partial x_1} = 2x_1 + 4x_2 + 1 = 0 \quad ① \\ \frac{\partial f}{\partial x_2} = 4x_1 + 2x_2 - 1 = 0 \quad ② \end{cases} \Rightarrow \begin{cases} x_1 = \frac{1}{2} \\ x_2 = -\frac{1}{2} \end{cases}$$

i.e. $(\frac{1}{2}, -\frac{1}{2})$ is the only critical point. ✓

(in next page)

2.

(a)

✓ let $g(x) = \frac{1}{2}\|Ax-b\|_2^2$

$$g(y) = \frac{1}{2}\|Ay-b\|_2^2 = \frac{1}{2}\|A(x+y-x)-b\|_2^2$$

$$= \frac{1}{2}\|Ax-b\|_2^2 + \langle Ax-b, Ay-Ax\rangle + \frac{1}{2}\|Ay-Ax\|_2^2$$

$$= g(x) + \langle Ax-b, A(y-x)\rangle + \frac{1}{2}\|A(y-x)\|_2^2$$

$$= g(x) + \langle A^T(Ax-b), y-x\rangle + \frac{1}{2}\|A(y-x)\|_2^2 \quad (*)$$

$\langle u, Av\rangle = u^TAv = (A^Tu)^Tv = \langle A^Tu, v\rangle$

$$\lim_{\|y-x\|_2\to 0} \frac{|g(y) - (g(x) + \langle A^T(Ax-b), y-x\rangle)|}{\|y-x\|_2}$$

$$\overset{(*)}{=} \lim_{\|y-x\|_2\to 0} \frac{\frac{1}{2}\|A(y-x)\|_2^2}{\|y-x\|_2}$$

by Cauchy-Schwarz inequality

$$\leq \lim_{\|y-x\|_2\to 0} \frac{\frac{1}{2}\|A\|_2^2 \cdot \|y-x\|_2^2}{\|y-x\|_2} \simeq 0$$

$$\Rightarrow \nabla g(x) = A^T(Ax-b)$$ ✓

✓ let $h(x) = \lambda\|x\|_2^2$

$$\nabla h(x) = \lambda\cdot(2x) = 2\lambda x$$

$$\text{So, } \nabla f(x) = \nabla(g(x)+h(x)) = \nabla g(x) + \nabla h(x)$$

$$= A^T(Ax-b) + 2\lambda x$$

(b) Assume the (i,j)-entry of X is x_{ij}, denote i-th entry of c as c_i

$$f(X) = b^TXc = \langle b, Xc\rangle = \sum_{i=1}^{n}\left(b_i \cdot \sum_{j=1}^{n} x_{ij}\cdot c_j\right)$$

$\sum_{j=1}^{n} x_{ij}\cdot c_j = (Xc)_i$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i x_{ij} c_j$$

$$\frac{\partial f(X)}{\partial x_{ij}} = b_i \cdot c_j \text{ , which is exactly the } (i,j)\text{ - entry of the matrix } bc^T$$

$$\text{i.e. } \frac{\partial f(X)}{\partial x_{ij}} = b_i c_j = (bc^T)_{ij}$$

$$\text{as } \frac{\partial f(x)}{\partial X} = \left[\frac{\partial f(X)}{\partial x_{ij}}\right],$$

$$\text{we have } \frac{\partial f(X)}{\partial X} = \left[\frac{\partial f(x)}{\partial x_{ij}}\right] = \left[(bc^T)_{ij}\right] = bc^T \checkmark$$

($A=[a_{ij}]$ means the $(i\text{-}j)$-entry of A is a_{ij})

(c) $f(X) = b^T X^T X c = (Xb)^T (Xc)$

$$= \langle Xb, Xc \rangle$$

$$= \sum_{i=1}^{n} \left(\underbrace{\left(\sum_{j=1}^{n} x_{ij} \cdot b_j \right)}_{(Xb)_i} \cdot \underbrace{\left(\sum_{j=1}^{n} x_{ij} \cdot c_j \right)}_{(Xc)_i} \right)$$

$$\text{Then } \frac{\partial f(x)}{\partial x_{ij}} = \frac{\partial (Xb)_i}{\partial x_{ij}} \cdot (Xc)_i + (Xb)_i \cdot \frac{\partial (Xc)_i}{\partial x_{ij}}$$

$$= \underbrace{b_j \cdot \left(\sum_{j=1}^{n} x_{ij} \cdot c_j \right)}_{\text{denote as } I_1} + \underbrace{c_j \cdot \left(\sum_{j=1}^{n} x_{ij} \cdot b_j \right)}_{\text{denote as } I_2}$$

$$= I_1 + I_2$$

Note that

$$\left(Xbc^T\right)_{ij} = \underbrace{\left(\sum_{j=1}^{n} x_{ij} \cdot b_j\right)}_{\searrow (Xb)_i} \cdot c_j \quad = I_1$$

$$\left(Xcb^T\right)_{ij} = \underbrace{\left(\sum_{j=1}^{n} x_{ij} \cdot c_j\right)}_{\searrow (Xc)_i} \cdot b_j \quad = I_2$$

Then we have $\dfrac{\partial f(X)}{\partial x_{ij}} = I_1 + I_2 = \left(Xbc^T\right)_{ij} + \left(Xcb^T\right)_{ij}$

$$= \left(Xbc^T + Xcb^T\right)_{ij}$$

$$= \left(X\left(bc^T + cb^T\right)\right)_{ij}$$

So $\dfrac{\partial f(X)}{\partial X} = \left[\dfrac{\partial f(X)}{\partial x_{ij}}\right] = \left[\left(X\left(bc^T + cb^T\right)\right)_{ij}\right]$

$$= X\left(bc^T + cb^T\right) \checkmark$$

3.

(a)

Proof:

Part 1: Prove: For $\forall\, b \in R^n$, b can be decomposed as

$$b = b_S + \sum_{i=1}^{N} c_i x_i$$

where $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in R^N$ and $b_S \in R^n$ satisfying $\langle b_S, x_i \rangle = 0$ for $\forall\, i = 1, \cdots, N$.

Proof: For simplicity, we prove only the case $N=1$.

Let $S = \{ v \in R^n \mid \langle v, x_1 \rangle = 0 \}$

Then S is a hyperplane and a subspace.

so b can be decomposed as

$$b = P_S b + (b - P_S b)$$

For $P_S b$: Since $P_S b$ is the projection,

$$\left.\begin{array}{l} b - P_S b \perp S \\ P_S b \in S \end{array}\right\} \Rightarrow \langle b - P_S b, P_S b \rangle = 0$$

For $b - P_S b$, By the explicit formular of projection,

$$b - P_S b = \frac{\langle x_1, b \rangle}{\|x_1\|^2} \cdot x_1 \equiv c_1 x_1 .$$

Define $P_S b = b_S$, then

$$b = b_S + c_1 x_1$$

and $\langle b - P_S b, P_S b \rangle = \langle c_1 x_1, b_S \rangle = 0$

$$\Rightarrow \langle x_1, b_S \rangle = 0. \quad \text{Proved.}$$

the conclusion can be generalized to $N = 2, 3, \cdots$ similarly.

Part 2: Then the function can be rewritten as:

$$\sum_{i=1}^{N} \left(\langle a, x_i \rangle - y_i \right)^2 + \lambda \|a\|_2^2$$

$$= \sum_{i=1}^{N} \left(\langle a_S + \sum_{j=1}^{N} c_j x_j, x_i \rangle - y_i \right)^2 + \lambda \left\| a_S + \sum_{j=1}^{N} c_j x_j \right\|_2^2$$

$$= \sum_{i=1}^{N} \left(\langle a_s, x_i \rangle + \langle \sum_{j=1}^{N} c_j x_j, x_i \rangle - y_i \right)^2$$

$$+ \lambda \left(\|a_s\|_2^2 + 2 \sum_{j=1}^{N} c_j \langle a_s, x_j \rangle + \sum_{j=1}^{N} \sum_{i=1}^{N} c_i c_j \langle x_i, x_j \rangle \right)$$

$\langle a_s, x_i \rangle = 0$ for $i = 1, \cdots, N$

$$= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} c_j \langle x_j, x_i \rangle - y_i \right)^2 + \lambda \sum_{j=1}^{N} \sum_{i=1}^{N} c_i c_j \langle x_i, x_j \rangle$$

$$+ \lambda \|a_s\|_2^2$$

Define $K = [\langle x_i, x_j \rangle]_{i,j=1}^{N}$

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{N} K_{ji} c_j - y_i \right)^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K_{ij} + \lambda \|a_s\|_2^2$$

$$= \sum_{i=1}^{N} \left((k^T c)_i - y_i \right)^2 + \lambda c^T K c + \lambda \|a_s\|_2^2$$

$$= \underbrace{\|K^T c - y\|_2^2 + \lambda c^T K c}_{\text{denote as } F(c)} + \underbrace{\lambda \|a_s\|_2^2}_{\text{denote as } G(a_s)}$$

Then,

the original optimization $\Leftrightarrow$

$$\begin{cases} \min\limits_{\substack{c \in R^N \\ a_s \in R^n}} F(c) + \lambda G(a_s) \\ \text{s.t. } \langle a_s, x_i \rangle = 0 \text{ for } i = 1, \cdots, N. \end{cases} \quad (\star)$$

$$\Leftrightarrow \min_{c \in R^N} F(c) \quad ①$$

and

$$\begin{cases} \min\limits_{a_s \in R^n} \lambda G(a_s) \\ \text{s.t. } \langle a_s, x_i \rangle = 0 \text{ for } i = 1, \cdots, N. \end{cases} \quad ②$$

To solve ②: Because $\lambda > 0$,

(in next page)

$$② \Longleftrightarrow \begin{cases} \min\limits_{a_s \in R^n} \|a_s\|_2^2 \\ s.t \quad \langle a_s, x_i \rangle = 0 \quad , \quad i = 1, \cdots, N \end{cases}$$

$$\Longleftrightarrow \boxed{a_s^* = 0}.$$

Let a^* be a solution of the original optimization, then

$$a^* = a_s^* + \sum_{i=1}^{N} c_i^* x_i$$

$$= \underbrace{\sum_{i=1}^{N} c_i^* x_i}_{\prime}$$

where $c^* = \begin{bmatrix} c_1^* \\ \vdots \\ c_N^* \end{bmatrix}$ is the solution of ①.

proved.

(b) In part (a) we have already obtained the re-expression in the process of proof, as (☆) above:

$$\min_{a \in R^n} \sum_{i=1}^{N} \left(\langle a, x_i \rangle - y_i \right)^2 + \lambda \|a\|_2^2$$

$$\Longleftrightarrow \min_{\substack{c \in R^N \\ a_s \in R^n \\ \langle a_s, x_i \rangle = 0, i=1,\cdots,N}} \sum_{i=1}^{N} \left(\|K^T c - y\|_2^2 + \lambda c^T K c \right) + \lambda \|a_s\|_2^2 \quad (\star)$$

where $K = \left[\langle x_i, x_j \rangle \right]_{i,j=1}^{N}$

As we have proved in (a):

To solve:

$$\min_{a_s \in R^n} \lambda \|a_s\|_2^2 \quad \text{s.t.} \quad \langle a_s, x_i \rangle = 0,\ i = 1, \cdots, N$$

we have a_s^* = the solution such that $\lambda \|a_s^*\|_2^2 = 0$ $(*-2)$

Thus

$(*) \overset{(*-2)}{\Longleftrightarrow}$

$$\min_{c \in R^N} \sum_{i=1}^{N} \left(\|K^T c - y\|_2^2 + \lambda c^T K c \right)$$

where $K = [\langle x_i, x_j \rangle]_{i,j=1}^{N}$.

is the final re-expression we want to find.

(in next page)

4. (a) Proof:

① First we get the gradient of $y = x^TAx$, where $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in R^n$, $A = [a_{ij}] \in R^{n\times n}$.

$$\nabla y = \frac{\partial y}{\partial x} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n} \right]^T,$$

$$y = x^TAx = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j.$$

then $$\frac{\partial y}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j\right)$$

$$= \sum_{i\neq k}\left(\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{n} a_{ij}x_ix_j\right)\right) + \sum_{j=1}^{n} a_{kj}x_kx_j$$

$$= \sum_{i\neq k}\left(\frac{\partial}{\partial x_k}\left(\sum_{j\neq k} a_{ij}x_ix_j\right) + \frac{\partial}{\partial x_k}a_{ik}x_ix_k\right) + \sum_{j\neq k}\frac{\partial}{\partial k}a_{kj}x_kx_j + \frac{\partial}{\partial k}a_{kk}x_k^2$$

(the first term is 0 since $i\neq k, j\neq k$)

$$= \sum_{i\neq k}\frac{\partial}{\partial x_k}a_{ik}x_ix_k + \sum_{j\neq k}\frac{\partial}{\partial k}a_{kj}x_kx_j + \frac{\partial}{\partial k}a_{kk}x_k^2$$

$$= \left(\sum_{i\neq k} a_{ik}x_i + a_{kk}x_k\right) + \left(\sum_{j\neq k} a_{kj}x_j + a_{kk}x_k\right)$$

$$= \sum_{i=1}^{n} a_{ik}x_i + \sum_{j=1}^{n} a_{kj}x_j$$

$$= (Ax)_k + (A^Tx)_k$$

(k-th entry of Ax)

Since A is symmetric, $A = A^T$

thus $$\frac{\partial y}{\partial x_k} = (2Ax)_k$$

thus $$\boxed{\Delta y = 2A\Delta x}$$

② Then we prove: if $A^{n\times n}$ is symmetric, then $\boxed{d^TAe = e^TAd}$ for $\forall e, d \in R^n$

As d^TAe is a scalar, $d^TAe = (d^TAe)^T$

$$= e^TA^Td$$

(as A is symmetric)

$$= e^TAd. \quad \text{proved.}$$

③ Then we prove: $f(x)$ is ③-1 differentiable and ③-2 convex.

3-1

$\nabla f(x) \overset{\text{as we have proved in ①}}{=\!=\!=} 2Ax + 2b$, so $f(x)$ is differentiable at $\forall x \in R^n$.

3-2

let $g(x) = x^TAx$ and $h(x) = 2b^Tx + C$

then $f(x) = g(x) + h(x)$.

for $h(x)$:

$h(y) = 2b^Ty + C$

$= 2b^T(x + y - x) + C$

$= 2b^Tx + 2b^T(y-x) + C$

$= h(x) + \langle \nabla f(x), y-x \rangle$, $\forall x, y \in R^n$.

Standard $\checkmark$ inner product $\langle \cdot, \cdot \rangle$ is induced norm on R^n.

so, $h(x)$ is (differentiable and) convex.

for $g(x)$:

$g(y) = y^TAy$

$= (x + (y-x))^T A (x + (y-x))$

$= (x^T + (y-x)^T) A (x + (y-x))$

$= x^TAx + \underline{x^TA(y-x)} + \underline{(y-x)^TAx} + (y-x)^TA(y-x)$

equal to each other as we have proved in ②.

$= x^TAx + \underline{2x^TA(y-x)} + (y-x)^TA(y-x)$

$= x^TAx + \langle 2Ax, y-x \rangle + \underline{(y-x)^TA(y-x)}$

≥ 0 as A is spsd. $\checkmark$

$\geq g(x) + \langle \nabla g(x), y-x \rangle$, $\forall x, y \in R^n$

so, $g(x)$ is (differentiable and) convex.

Altogether, for $f(x)$:

$$\begin{cases} h(y) = h(x) + \langle \nabla h(x), y-x \rangle & (*-1) \\ g(y) \geq g(x) + \langle \nabla g(x), y-x \rangle & (*-2) \end{cases}$$

$(*-1) + (*-2)$, we have:

(in next page)

$$h(y) + g(y) \geq h(x) + g(x) + \langle \nabla h(x) + \nabla g(x),\ y - x \rangle$$

i.e. $$f(y) \geq f(x) + \langle \nabla f(x),\ y - x \rangle$$ ✓

i.e $f(x)$ is (differentiable and) convex.

④ According to the theorem we discussed in the class,

since $f(x)$ is differentiable and convex,

the global minimizer $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x)$

$$\Updownarrow$$

$$\nabla f(x^{(*)}) = 0$$ ✓

$$\nabla f(x) = 2Ax + 2b = 0 \iff Ax = -b$$

i.e. $x = x^{(*)} \iff Ax = -b$. proved. 囡

(in next page)

4(b).

Notations: $Col(A) = \{ Ay \mid \forall y \in R^n \}$ (column space of $A_{n\times n}$)

$Nul(A) = \{ x \mid Ax = 0 \}$

$Nul(A)^\perp = \{ y \mid y \perp Nul(A)$ i.e. for $\forall x \in Nul(A)$, $\langle x, y \rangle = 0 \}$

Part 0: Prove: $Nul(A^T)^\perp = Col(A)$ (☆)

Proof: $\forall x \in R^n$, $\forall w \in Nul(A^T)$

$\langle \underbrace{Ax}_{\in Col(A)}, w \rangle = x^T \underbrace{A^T w}_{=0 \text{ since } w \in Nul(A^T)} = 0$

thus $Nul(A^T)^\perp = Col(A)$

Part 1: Prove: f is bounded below over R^n $\Rightarrow$ $b \in \{Ay : y \in R^n\}$. (i.e.) $b \in Col(A)$

Proof: We prove it by proving its converse-negative proposition.

We assume $b \notin Col(A)$, according to (☆).

we have $b \notin Nul(A^T)^\perp$

i.e. for $\forall x \in Nul(A^T)$, $\langle b, x \rangle \neq 0$ ✓

since $\exists b \notin Col(A)$ (as we have assumed), we can infer that $\dim(A) < n$.

$\dim(A) < n \Rightarrow \dim(Col(A)) < n \overset{☆}{\Longleftrightarrow} \dim\left(Nul(A^T)^\perp\right) < n$

$\Rightarrow \dim\left(Nul(A^T)\right) = n - \dim\left(Nul(A^T)^\perp\right) > 0$

so there must be non-zero vector in $Nul(A^T)$

thus we can find $x_1 \in Nul(A^T)$, $x_1 \neq 0$

such that $\langle b, x_1 \rangle = b^T x_1 \neq 0$.

Since $x_1 \in Nul(A^T)$, we have $A^T x_1 \overset{A^T = A \text{ as } A \text{ is spsd}}{=\!=} A x_1 = 0$

Then $f(\underbrace{-t\cdot sgn(b^Tx_1)\cdot x_1}_{\text{denote as } y})$

Substitute x with $x = -t\cdot sgn(b^Tx_1)\cdot x_1$ and then we finally get:

$$= -t\cdot sgn(b^Tx_1)\cdot y\underbrace{Ax_1}_{=0} + 2b^Ty + c$$

$$= -2t\cdot sgn(b^Tx_1)\cdot b^Tx_1 + c \overset{\text{denote as}}{=\!=} F$$

where $sgn(x) = \begin{cases} 1, & x>0 \\ -1, & x<0 \end{cases}$

$F \to -\infty$ as $t \to +\infty$, so we proved:

$b \notin \{Ay : y \in R^n\} \Rightarrow f$ is never bounded below over R^n"

so we can conclude that:

f is bounded below over $R^n \Rightarrow b \in \{Ay : y \in R^n\}$ ✓

Part 2: Prove: $b \in \{Ay : y \in R^n\} \Rightarrow f$ is bounded below over R^n.

Proof: Since $b \in \{Ay : y \in R^n\}$

$\exists\, y \in R^n$ such that $b = Ay$ (y is constant)

then $f(x) = x^TAx + 2b^Tx + c = x^TAx + 2x^TAy + c$

($x^TAy = y^TAx$ since A is symmetric) $= x^TAx + x^TAy + y^TAx + c$

$= x^TAx + x^TAy + y^TAx + y^TAy - y^TAy + c$

$= (x+y)^TA(x+y) - y^TAy + c$

≥ 0 since A is spsd.

$\geq -y^TAy + c$ ✓

thus, $-y^TAy+C$ is the lower boundary of f.

i.e. we have proved:

$b \in \{Ay : y \in R^n\} \Rightarrow f$ is bounded below over R^n.

Altogether,

$b \in \{Ay : y \in R^n\} \Longleftrightarrow f$ is bounded below over R^n.

(in next page)

5(a).

let $g(x) = \log(x), \quad x > 0$

$$h(x) = \sum_{i=1}^{m} \exp(a_i^T x + b_i) \quad , \quad x \in R^n.$$

$$g'(x) = \frac{1}{x} .$$

$$\nabla h(x) = \sum_{i=1}^{m} \nabla\left(\exp(a_i^T x + b_i)\right)$$

$$= \sum_{i=1}^{m} \left[\exp(a_i^T x + b_i) \cdot \nabla(a_i^T x + b_i)\right]$$

$$= \sum_{i=1}^{m} \left[\exp(a_i^T x + b_i) \cdot a_i\right]$$

thus $\nabla f(x) = \nabla(g(h(x)))$

$$= g'(h(x)) \cdot \nabla h(x)$$

$$= \frac{\sum_{i=1}^{m} \left[\exp(a_i^T x + b_i) \cdot a_i\right]}{\sum_{i=1}^{m} \exp(a_i^T x + b_i)}$$ ✓

(in next page)

5(b) Yes, it will.

Proof: [general idea: First we prove $f(x)$ is convex, then we further conclude the convergence in GD process.]

[note: $f(x) = \log(g(x))$, $x \in R^n$, $g(x): R^n \to R$, $g(x) > 0$ for $\forall x \in R^n$. if $f(x)$ is convex, we call $g(x)$ is log-convex.]

Part 1: Prove: for $\forall\, g(x): R^n \to R$ which satisfies $g(x) > 0$ for $\forall x \in R^n$, we have:

$g(x)$ is log-convex $\iff g(\theta x + (1-\theta)y) \le [g(x)]^{\theta}[g(y)]^{1-\theta}$

$\forall\, \theta \in [0,1]$, $\forall x, y \in R^n$.

proof: for $\forall\, g(x): R^n \to R$ and $g(x) > 0$ for $\forall x \in R^n$,

$g(x)$ is log-convex

$\iff f(x) = \log(g(x))$ is convex

by the definition of convexity

$\iff f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$, $\forall x, y \in R^n, \forall \theta \in [0,1]$

substitute f with log(g)

$\iff \log\big(g(\theta x + (1-\theta)y)\big) \le \theta \log(g(x)) + (1-\theta)\log(g(y))$, $\forall x, y \in R^n, \forall \theta \in [0,1]$ (*)

take the logarithm on both sides (backward)

$\iff\!\!\Rightarrow g(\theta x + (1-\theta)y) \le \exp\Big(\theta \log(g(x)) + (1-\theta)\log(g(y))\Big)$, $\forall x, y \in R^n, \forall \theta \in [0,1]$

take the exponential on both sides (forward)

$\iff g(\theta x + (1-\theta)y) \le \exp\big(\theta \log(g(x))\big) \cdot \exp\Big((1-\theta)\log(g(y))\Big)$, $\forall x, y \in R^n, \forall \theta \in [0,1]$

$\iff g(\theta x + (1-\theta)y) \le [g(x)]^{\theta} \cdot [g(y)]^{1-\theta}$, $\forall x, y \in R^n, \forall \theta \in [0,1]$.

proved.

Part 2: Prove: $h(x) = \exp(a^T x + b_i)$, $a, x \in R^n, b_i \in R$, is log-convex

Proof: $h(x) > 0$ for $\forall x \in R^n$ since exp functions are always positive.

$$h(\theta x + (1-\theta)y) = \exp\big(a^T(\theta x + (1-\theta)y) + b_i\big)$$

$$= \exp\big(\theta a^T x + (1-\theta)\, a^T y + b_i\big)$$

(in next page)

$$= \exp\left(\theta(a^T x + b_i) + (1-\theta)(a^T y + b_i)\right)$$

$$= \exp(\theta(a^T x + b_i)) \cdot \exp\left((1-\theta)(a^T y + b_i)\right)$$

$$= \left[\exp(a^T x + b_i)\right]^{\theta} \cdot \left[\exp(a^T y + b_i)\right]^{1-\theta}$$

$$= [h(x)]^{\theta} \cdot [h(y)]^{1-\theta} \leq [h(x)]^{\theta} \cdot [h(y)]^{1-\theta},$$

for $\forall x, y \in R^n$, $\forall \theta \in [0,1]$

According to Part1 → So $h(x)$ is log-convex. proved.

Part 3: Prove: if $a>0$, $b>0$, $\alpha, \beta \in [0,1]$ and $\alpha+\beta=1$ then $a^{\alpha} b^{\beta} \leq \alpha a + \beta b$.

Proof: First we prove: If $A>0$, $B>0$, $\frac{1}{p}+\frac{1}{q}=1$, $p>1$, $q>1$ then $AB \leq \frac{A^p}{p} + \frac{B^q}{q}$. (1)

th-1

Let $t = \frac{A^p}{B^q}$, $\varphi = \frac{1}{p}$,

then (1) $\Longleftrightarrow$ $\left(\frac{A^p}{B^q} \cdot B^q\right)^{\frac{1}{p}} \cdot \left(B^q\right)^{1-\frac{1}{p}} \leq \frac{1}{p}\left(\frac{A^p}{B^q} \cdot B^q\right) + \left(1-\frac{1}{p}\right) \cdot B^q$

$\Longleftrightarrow$ $(t \cdot B^q)^{\varphi} \cdot (B^q)^{1-\varphi} \leq \varphi(t \cdot B^q) + (1-\varphi) B^q$

$\Longleftrightarrow$ $t^{\varphi} \cdot B^q \leq (\varphi t + 1 - \varphi) B^q$

$\Longleftrightarrow$ $t^{\varphi} \leq \varphi t + 1 - \varphi$ (2)

Let $d(t) = t^{\varphi} - \varphi t$, thus $d'(t) = \varphi t^{\varphi-1} - \varphi = \varphi(t^{\varphi-1} - 1)$.

Since $p>1$, then $0<\varphi<1$, i.e, in $d'(t)$ the exponent of t is negative

so $d'(t)$ is monotonically decreasing in $t \in (0, +\infty)$,

And since $d'(t)$ is differentiable at $\forall t>0$, $(d''(t) = \varphi(\varphi-1)t^{\varphi-2})$, we have:

$\begin{cases} \text{when } t \in (0,1), & d'(t) > d'(1) = 0, \text{ thus } d(t) \text{ is monotonically increasing.} \\ \text{when } t > 1, & d'(t) < d'(1) = 0, \text{ thus } d(t) \text{ is monotonically decreasing.} \end{cases}$

Therefore, $t^{\varphi} - \varphi t = d(t) \leq d(1) = 1 - \varphi \Longleftrightarrow (2) \Longleftrightarrow (1)$

Now, we've proved th-1.

We reformulate th-1 by some transformation:

Let $\alpha = \frac{1}{p}$, $\beta = \frac{1}{q}$, $A^p = a$, $B^q = b$,

(thus $A>0$, $B>0$, $\alpha, \beta \in (0,1)$ and $\alpha+\beta=1$)

then $AB \le \frac{A^p}{p} + \frac{B^q}{q} \iff a^\alpha b^\beta \le \alpha a + \beta b$ (3)

There are still conditions with boundary values.

when $\alpha=1, \beta=0$, $a^\alpha b^\beta = a = \alpha a+\beta b$

when $\alpha=0, \beta=1$, $a^\alpha b^\beta = b = \alpha a+\beta b$

Altogether, we have proved:

When $a>0$, $b>0$, $\alpha,\beta \in [0,1]$ and $\alpha+\beta=1$
then $a^\alpha b^\beta \le \alpha a+\beta b$.

✓

Part 4: Prove: Assume $H_1(x), H_2(x): R^n \to R$, $H_1(x), H_2(x) > 0$ for $\forall x \in R^n$,
If $H_1(x)$ and $H_2(x)$ are both log-convex,
then $H(x) = H_1(x) + H_2(x)$ is also log-convex. (*-0)

Proof: Since $H_1(x)$, $H_2(x)$ are log-convex, as we proved in Part 1,

$$\begin{cases} H_1(\theta x+(1-\theta)y) \le [H_1(x)]^\theta [H_1(y)]^{1-\theta}, \ \forall x,y \in R^n, \ \forall \theta \in [0,1] & (4) \\ H_2(\theta x+(1-\theta)y) \le [H_2(x)]^\theta [H_2(y)]^{1-\theta}, \ \forall x,y \in R^n, \ \forall \theta \in [0,1] & (5) \end{cases}$$

(4)+(5), we have

$$H(\theta x+(1-\theta)y) = H_1(\theta x+(1-\theta)y) + H_2(\theta x+(1-\theta)y)$$
$$\le [H_1(x)]^\theta [H_1(y)]^{1-\theta} + [H_2(x)]^\theta [H_2(y)]^{1-\theta}$$

Notice that what we want to prove is: $H(x)$ is log-convex, i.e.,

$$H(\theta x+(1-\theta)y) \le [H(x)]^\theta [H(y)]^{1-\theta} = [H_1(x)+H_2(x)]^\theta [H_1(y)+H_2(y)]^{1-\theta}$$

So to prove (*-0), we just need to prove (*-1) as below

✓

$$[H_1(x)]^\theta [H_1(y)]^{1-\theta} + [H_2(x)]^\theta [H_2(y)]^{1-\theta} \le [H_1(x)+H_2(x)]^\theta [H_1(y)+H_2(y)]^{1-\theta} \quad (*-1)$$

let $a = H_1(x)$, $b = H_1(y)$, $c = H_2(x)$, $d = H_2(y)$
then to prove (*-1) becomes to prove (*-2) as below:

$$a^\theta b^{1-\theta} + c^\theta d^{1-\theta} \le (a+c)^\theta (b+d)^{1-\theta} \quad (*-2)$$

divide $\underbrace{(a+c)^\theta (b+d)^{1-\theta}}_{>0}$ on both side, we have

$$\frac{a^\theta b^{1-\theta} + c^\theta d^{1-\theta}}{(a+c)^\theta (b+d)^{1-\theta}} \le 1 \qquad (*-3)$$

Since $\begin{cases} a = H_1(x) \\ c = H_2(x) \end{cases}$, $\begin{cases} b = H_1(y) \\ d = H_2(y) \end{cases}$ there is no relationship between the value of a and c (or: b and d),

→ thus, $a+c>0$ can be arbitrary

similarly, $b+d>0$ can be arbitrary.

→ Thus, we assume that $\underbrace{a+c=k_1}$, $\underbrace{b+d=k_2}$, $k_1, k_2 > 0$ are both arbitrary.

and we set $a_1 = \frac{a}{k_1}$, $c_1 = \frac{c}{k_1}$, such that $a_1 + c_1 = 1$

similarly we set $b_1 = \frac{b}{k_2}$, $d_1 = \frac{d}{k_2}$, such that $b_1 + d_1 = 1$

So the LHS of (*-3) can be transformed as below:

$$\frac{a^\theta b^{1-\theta} + c^\theta d^{1-\theta}}{(a+c)^\theta (b+d)^{(1-\theta)}} = \frac{(k_1 a_1)^\theta (k_2 b_1)^{1-\theta} + (k_1 c_1)^\theta (k_2 d_1)^{1-\theta}}{k_1^\theta (a_1+c_1)^\theta \, k_2^{1-\theta} (b_1+d_1)^{1-\theta}}$$

$$= \frac{\left(k_1^\theta k_2^{1-\theta}\right) a_1^\theta b_1^{1-\theta} + \left(k_1^\theta k_2^{1-\theta}\right) c_1^\theta d_1^{1-\theta}}{\left(k_1^\theta k_2^{1-\theta}\right) (a_1+c_1)^\theta (b_1+d_1)^{1-\theta}}$$

$$= \frac{a_1^\theta b_1^{1-\theta} + c_1^\theta d_1^{1-\theta}}{(a_1+c_1)^\theta (b_1+d_1)^{1-\theta}} \qquad (6)$$

According to (6), we can reformulate (*-3) as:

$$\frac{a_1^\theta b_1^{1-\theta} + c_1^\theta d_1^{1-\theta}}{(a_1+c_1)^\theta (b_1+d_1)^{1-\theta}} \le 1 \iff a_1^\theta b_1^{1-\theta} + c_1^\theta d_1^{1-\theta} \le 1,$$

where $a_1 + c_1 = b_1 + d_1 = 1$

(*-4)

Thus, to prove (*-0), we just need to prove (*-4). ☆

$a_1, b_1, c_1, d_1 > 0$, $\theta \in [0,1]$

As we have proved in Part 3,

$$a_1^\theta + b_1^{1-\theta} \le \theta a_1 + (1-\theta) b_1 \qquad (7)$$

(in next page)

$$c_1^{\theta} d_1^{1-\theta} \leq \theta c_1 + (1-\theta) d_1 \quad (8)$$

(7)+(8), we have:

$$a_1^{\theta} b_1^{1-\theta} + c_1^{\theta} d_1^{1-\theta} \leq \theta a_1 + (1-\theta) b_1 + \theta c_1 + (1-\theta) d_1$$

$$= \theta(a_1 + c_1) + (1-\theta)(b_1 + d_1)$$

$$= \theta + 1 - \theta$$

$$= 1$$

Thus we proved (*-4),

i.e, we proved (*-0).

Part 5: Prove: $j(x) = \sum_{i=1}^{m} \exp(a_i^T x + b_i)$ is log-convex

$a_i, x \in R^n, b_i \in R.$

let $h_i(x) = \exp(a_i^T x + b_i)$, $i = 1, 2, \cdots, m$,
then $j(x) = \sum_{i=1}^{m} h_i(x)$.

As we have proved in Part 2,

$h_i(x)$ is log-convex, $i = 1, 2, \cdots, m$

Then as we have proved in Part 4,

$h_1(x) + h_2(x)$ is log-convex.

Then $(h_1(x) + h_2(x)) + h_3(x)$ is log-convex

since $h_1(x) + h_2(x)$ is convex (log-) and $h_3(x)$ is convex (log-).

...

Thus, inductively, we can finally conclude that

$$j(x) = \sum_{i=1}^{m} h_i(x) \text{ is log-convex.}$$

Part 6: Get the conclusion.

Since $j(x)$ is log-convex.

$f(x) = \log(j(x))$ is convex (condition-1)

$= \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right)$

According to 5(a), we know $f(x)$ is differentiable. (condition-2)

Some interpretion for (condition-2):

And since $a_i^T x + b_i$, $x \in R^n$, $\exp(x)$, $x \in R$, $\log(x)$, and sum-function

are all differentiable in their respective definition domains.

so $f(x)$ which is composed of these differentiable function

is also differentiable. (condition-2')

by { condition-1, condition-2 }, we have

$$x^{(*)} = \underset{x \in R^n}{\operatorname{argmin}} f(x) \Leftrightarrow \nabla f(x^{(*)}) = 0.$$

If we use gradient descent to solve $\min_{x \in R^n} f(x)$,

and GD converges, then

we have

$$\lim_{k \to +\infty} \nabla f(x^{(k)}) = 0 \iff \lim_{k \to +\infty} x^{(k)} = \underset{x \in R^n}{\operatorname{argmin}} f(x)$$

i.e. It will converge to the global minimizer.

The minimizer may not exist. →