Homework 1

Zhaoheng Zheng
EECS 542 - Machine Learning

September 17, 2017

1) (a) (i) *Proof.* Let symmetric matrix $A \in \mathbb{R}^{2\times 2}$,

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

The inverse of A is,

$$A^{-1} = \begin{bmatrix} \frac{d}{ad-b^2} & \frac{-b}{ad-b^2} \\ \frac{-b}{ad-b^2} & \frac{a}{ad-b^2} \end{bmatrix}$$

Therefore $(A^{-1})^T = A^{-1}$, A^{-1} is a symmetric matrix, the statement is *True*. □

(ii) *Proof.* We assume that matrix $A \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is orthogonal

Hence

$$AA^{-1} = AA^T = \begin{bmatrix} a^2+b^2 & ac+bd \\ ac+bd & c^2+d^2 \end{bmatrix} = I$$

Then we have equations

$$\begin{aligned} a^2+b^2 &= 1 \\ c^2+d^2 &= 1 \\ ac+bd &= 0 \end{aligned}$$

We can see that point $(a,b),(c,d)$ are two points on unit circle. So $\exists \theta, \theta' \in \mathbb{R}$ that satisfy

$$a = cos\theta, b = sin\theta$$
$$c = cos\theta', d = sin\theta'$$

Hence

$$\begin{aligned} ac+bd &= cos\theta cos\theta' + sin\theta sin\theta' \\ &= cos(\theta' - \theta) \\ &= 0 \end{aligned}$$

From the equation above we have

$$\theta' = \theta + (k+\frac{1}{2})\pi, k \in \mathbb{Z}$$

When k is even,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ cos\theta' & sin\theta' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} = \begin{bmatrix} cos(-\theta) & -sin(-\theta) \\ sin(-\theta) & cos(-\theta) \end{bmatrix}$$

When k is odd,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ cos\theta' & sin\theta' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ sin\theta & -cos\theta \end{bmatrix}$$

Therefore the statement is *True*. □

(iii) *Disproof.* Assume that there exists matrix $C \in \mathbb{R}^{3\times 3}$ satisfies

$$A = CC^T$$

Then for any vector $X \in \mathbb{R}^3$,

$$\begin{aligned} XAX^T &= XCC^TX^T \\ &= (XC)(XC)^T \\ &= \langle XC, XC \rangle \geq 0 \end{aligned}$$

But if we let

$$X = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

Then

$$XAX^T = -8 < 0$$

contradicts the inequality $XAX^T \geq 0$
Therefore the statement if *False*. □

2) (a) (i) *Proof.*

$$\begin{aligned} p_X(x) &= \int_y p(x,y)dy \\ &= \int_y p_{X|Y}(x|y)p_Y(y)dy \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}[X] &= \int_x xp_X(x)dx \\ &= \int_x x \int_y p_{X|Y}(x|y)p_Y(y)dydx \\ &= \int_y p_Y(y) \int_x xp_{X|Y}(x|y)dxdy \\ &= \int_y p_Y(y)\mathbb{E}_X[X|Y]dy \\ &= \mathbb{E}_Y[\mathbb{E}_X[X|Y]] \end{aligned}$$

$\square$

(ii) *Proof.*

$$\begin{aligned}\mathbb{E}[I[X \in \mathcal{C}]] &= \int_x I[x \in \mathcal{C}] p_X(x) dx \\ &= \int_{x \in \mathcal{C}} p_X(x) dx \\ &= P(X \in \mathcal{C})\end{aligned}$$

$\square$

(iii) *Proof.* We first prove $var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$:

$$\begin{aligned}\text{var}[X] &= \int_x (x - \mathbb{E}[X])^2 p_X(x) dx \\ &= \int_x (x^2 - 2x\mathbb{E}[X] + [\mathbb{E}[X]]^2) p_X(x) dx \\ &= \mathbb{E}(X^2) - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}[X])^2\end{aligned}$$

Hence, we obtain

$$\mathbb{E}_Y[\text{var}_X[X|Y]] = \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2]] = \mathbb{E}[X^2] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2]$$

$$\text{var}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]^2] - \mathbb{E}_Y[\mathbb{E}_X[X|Y]]^2 = \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - (\mathbb{E}[X])^2$$

Therefore $\mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{var}[X]$. $\square$

(iv) *Proof.* Since X and Y are independent, we obtain $p(x, y) = p_x(x)p_y(y)$. Therefore,

$$\begin{aligned}\mathbb{E}[XY] &= \int_x \int_y xyp(x, y) dy dx \\ &= \int_x \int_y x p_X(x) y p_Y(y) dy dx \\ &= \int_x x p_X(x) \left[\int_y y p_Y(y) dy \right] dx \\ &= \mathbb{E}[Y] \int_x x p_X(x) dx \\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

$\square$

(v) *Proof.* Since X and Y takes value in $\{0, 1\}$, we obtain

$$\mathbb{E}[X] = P(X = 1), \quad \mathbb{E}[Y] = P(Y = 1)$$

Then we can get

$$\mathbb{E}[XY] = P(X = 1, Y = 1) = \mathbb{E}[X]\mathbb{E}[Y] = P(X = 1)P(Y = 1)$$

Hence

$$P(X=1,Y=0)=P(X=1)-P(X=1,Y=1)=P(X=1)P(Y=0)$$
$$P(X=0,Y=1)=P(Y=1)-P(X=1,Y=1)=P(X=0)P(Y=1)$$
$$P(X=0,Y=0)=P(X=0)-P(X=0,Y=1)=P(X=0)P(Y=0)$$

Therefore $P(X,Y)=P(X)P(Y), X, Y$ are independent. □

(b) (i) $P(H=h,D=d)\leq P(H=h)$

Proof. We can obtain that

$$P(H=h)=\sum_d P(H=h,D=d)$$

For any $d, P(H=h,D=d)\geq 0$,
Therefore $P(H=h)=\sum_d P(H=h,D=d)\geq P(H=h,D=d)$. □

(ii) It depends.
Since $P(H=h|D=d)=\frac{P(H=h,D=d)}{P(D=d)}$ and we can't decide the value of $P(D=d)\in(0,1)$, it depends.

(iii) $P(H=h|D=d)\geq P(D=d|H=h)P(H=h)$,

Proof.

$$P(H=h|D=d)=\frac{P(H=h,D=d)}{P(D=d)}$$

Therefore

$$P(D=d|H=h)P(H=h)=P(H=h,D=d)\leq P(H=h|D=d)$$

□

3) (a) *Proof.* Since $U^TU=UU^T=I$, we can obtain that for any column vector $\mathbf{u}_i$. So

$$\mathbf{u}_i^T\mathbf{u}_i=1$$

We first assume that matrix A is PSD, then for each eigenvalue λ_i , we obtain

$$\mathbf{u}_i^TA\mathbf{u}_i=\mathbf{u}_i^T\lambda_i\mathbf{u}_i=\lambda_i\mathbf{u}_i^T\mathbf{u}_i=\lambda_i\geq 0$$

Then we assume for each i, $\lambda_i\geq 0$. For all $\mathbf{x}\in\mathbb{R}^d$,

$$\begin{aligned}\mathbf{x}^TA\mathbf{x}&=\mathbf{x}^TU\Lambda U^T\mathbf{x}\\&=\mathbf{x}^T\sum_{i=1}^d\lambda_i\mathbf{u}_i\mathbf{u}_i^T\mathbf{x}\\&=\sum_{i=1}^d\lambda_i(\mathbf{x}^T\mathbf{u}_i)(\mathbf{x}^T\mathbf{u}_i)^T\\&=\sum_{i=1}^d\lambda_i(\mathbf{x}^T\mathbf{u}_i)^2\end{aligned}$$

Since for each $i, \lambda_i \geq 0$. Hence for each $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)^2 \geq 0$$

Therefore A is PSD iff $\lambda_i \geq 0$ for each i. □

(b) *Proof.* We first assume that matrix A is PD, then for each eigenvalue λ_i , we obtain

$$\mathbf{u}_i^T A\mathbf{u}_i = \mathbf{u}_i^T \lambda_i \mathbf{u}_i = \lambda_i \mathbf{u}_i^T \mathbf{u}_i = \lambda_i > 0$$

Then we assume for each i, $\lambda_i > 0$. For all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}\mathbf{x}^T A\mathbf{x} &= \mathbf{x}^T U\Lambda U^T \mathbf{x} \\ &= \mathbf{x}^T \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \mathbf{x} \\ &= \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)(\mathbf{x}^T \mathbf{u}_i)^T \\ &= \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)^2\end{aligned}$$

Now we prove that there is no $\mathbf{x} \neq \mathbf{0}, \mathbf{x}^T \mathbf{u}_i = 0$ for each i.
Assume that for each i, we have $\mathbf{x} \neq \mathbf{0}$ that $\mathbf{x}^T \mathbf{u}_i = 0$, we can let matrix $X \in \mathbb{R}^d$ be

$$X = \begin{bmatrix}\mathbf{x} \cdots \mathbf{x}\end{bmatrix}$$

Then

$$X^T U U^T = \begin{bmatrix}\mathbf{x}^T \\ \vdots \\ \mathbf{x}^T\end{bmatrix} \begin{bmatrix}\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_d\end{bmatrix} U^T = \mathbf{0} \cdot U^T = \mathbf{0}$$

However,

$$X^T U U^T = X^T I = X^T \neq \mathbf{0}$$

which contradicts the equation above. Hence there is no $\mathbf{x} \neq \mathbf{0}, \mathbf{x}^T \mathbf{u}_i = 0$ for each i.
Therefore

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)^2 > 0$$

implies A is PD. □

4) (a) *Proof.*

$$\begin{aligned}f(t\mathbf{x} + (1-t)\mathbf{y}) &= \mathbf{a}^T (t\mathbf{x} + (1-t)\mathbf{y}) + b \\ &= t(\mathbf{a}^T x + b) + (1-t)(\mathbf{a}^T y + b) \\ &= tf(\mathbf{x}) + (1-t)f(\mathbf{y})\end{aligned}$$

□

So affine function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ is convex. And

$$\begin{aligned}
-f(t\mathbf{x} + (1-t)\mathbf{y}) &= -(\mathbf{a}^T(t\mathbf{x} + (1-t)\mathbf{y}) + b) \\
&= -t(\mathbf{a}^T x + b) - (1-t)(\mathbf{a}^T y + b) \\
&= t(-f(\mathbf{x})) + (1-t)(-f(\mathbf{y}))
\end{aligned}$$

Therefore $f(\mathbf{x})$ is convex and concave.
Since for all $\mathbf{x}, \mathbf{y}$ in the domain of f,

$$f(t\mathbf{x} + (1-t)\mathbf{y}) = tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

$f(\mathbf{x})$ is not strictly convex.

(b) *Proof.* Assume that f has more than one global minimizers, we denote two of them as $\mathbf{x}_1, \mathbf{x}_2$.
For any $t \in [0,1]$, we let $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_2$. Since f is strictly convex, we can obtain that inequality:

$$f(\mathbf{x}_t) = f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) < tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2) = f(\mathbf{x}_1) = f(\mathbf{x}_2)$$

which contradicts the statement that $\mathbf{x}_1, \mathbf{x}_2$ are global minimizers.
Therefore, if f is strictly convex, then f has at most one global minimizer. □

(c) *Proof.* Since $\mathbf{x}^*$ is a local minimizer, we can obtain

$$\nabla f(\mathbf{x}^*) = 0$$

For all $\mathbf{x}$ in the neighborhood of $\mathbf{x}^*$,

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{x}^*) &= \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + \frac{1}{2}\langle \mathbf{x} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \rangle + o(||\mathbf{x} - \mathbf{x}^*||^2) \\
&= \frac{1}{2}\langle \mathbf{x} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \rangle + o(||\mathbf{x} - \mathbf{x}^*||^2) \geq 0
\end{aligned}$$

Then

$$\begin{aligned}
\lim_{\mathbf{x} \to \mathbf{x}^*} \frac{f(\mathbf{x}) - f(\mathbf{x}^*)}{||\mathbf{x} - \mathbf{x}^*||^2} &= \lim_{\mathbf{x} - \mathbf{x}^*} \left[\frac{1}{2} \frac{\langle \mathbf{x} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \rangle}{||\mathbf{x} - \mathbf{x}^*||^2} + \frac{o(||\mathbf{x} - \mathbf{x}^*||^2)}{||\mathbf{x} - \mathbf{x}^*||^2} \right] \\
&= \lim_{\mathbf{x} - \mathbf{x}^*} \left[\frac{1}{2} \frac{\langle \mathbf{x} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \rangle}{||\mathbf{x} - \mathbf{x}^*||^2} \right] \geq 0
\end{aligned}$$

Hence $\langle \mathbf{x} - \mathbf{x}^*, \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \rangle \geq 0$.
So for all direction $\mathbf{h}$,

$$\langle \mathbf{h}, \nabla^2 f(\mathbf{x}^*) h \rangle \geq 0$$

Therefore $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite. □

(d) *Proof.* We first assume that $\nabla^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$. Denote $g(t) = f(t\mathbf{x} + (1-y)\mathbf{y})$, we can obtain

$$g'(t) = (\mathbf{x} - \mathbf{y})^T \nabla f(t\mathbf{x} + (1-t)\mathbf{y})$$

$$g''(t) = (\mathbf{x} - \mathbf{y})^T \nabla^2 f(t\mathbf{x} + (1-t)\mathbf{y})(\mathbf{x} - \mathbf{y})$$

Since $\nabla^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$, we can get that for all t, $g''(t) \geq 0$.

Then according to Taylor series, we obtain

$$g(0) = g(t) + g'(t)(-t) + \frac{1}{2}g''(t - \theta t)t^2 \geq g(t) + g'(t)(-t)$$

$$g(1) = g(t) + g'(t)(1-t) + \frac{1}{2}g''(t + \theta(1-t))(1-t)^2 \geq g(t) + g'(t)(1-t)$$

Thus for $t \in [0, 1]$,

$$(1-t)g(0) + tg(1) \geq (1-t)g(t) + g'(t)(1-t)(-t) + tg(t) + g'(t)(1-t)t = g(t)$$

which implies

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

Hence $f(\mathbf{x})$ is convex.

Now we assume $f(\mathbf{x})$ is convex, by definition, for any $t \in [0, 1]$

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

We can rewrite the inequality into

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq t[f(\mathbf{x}) - f(\mathbf{y})] + f(\mathbf{y})$$

Then for any $t \in (0, 1]$,

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \frac{f(t\mathbf{x} + (1-t)\mathbf{y}) - f(\mathbf{y})}{t}$$

When $t \to 0$, we get

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

implies

$$\frac{1}{2}\frac{\langle \mathbf{x} - \mathbf{y}, \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y}) \rangle}{||\mathbf{x} - \mathbf{y}||^2} + \frac{o(||\mathbf{x} - \mathbf{y}||^2)}{||\mathbf{x} - \mathbf{y}||^2}$$

Limit $\mathbf{x} \to \mathbf{y}$, $\langle \mathbf{x} - \mathbf{y}, \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y}) \rangle \geq 0$ for all $\mathbf{y} \in \mathbb{R}^d$.

Therefore, $\nabla^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$ □

(e) *Solution.* The first derivative $\nabla f(\mathbf{x})$ is $A\mathbf{x} + \mathbf{b}$, while the Hessian is A.

When A is a positive semi-definite matrix, f is convex.

When A is positive definite, f is strictly convex. □

5) (a) The result and ratio for k-approximation are shown below:

(i) $k = 2, \frac{||X-\tilde{X}||_F}{||X||_F} = 0.281484,$

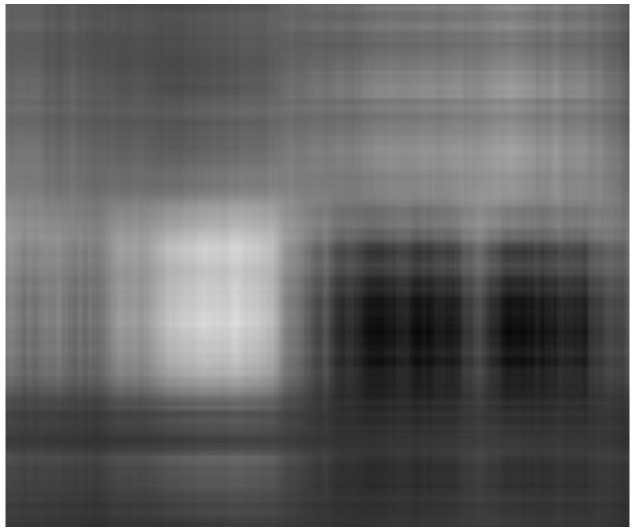

Figure 1: Approximation image for k=2

(ii) $k = 10, \frac{||X-\tilde{X}||_F}{||X||_F} = 0.158739,$

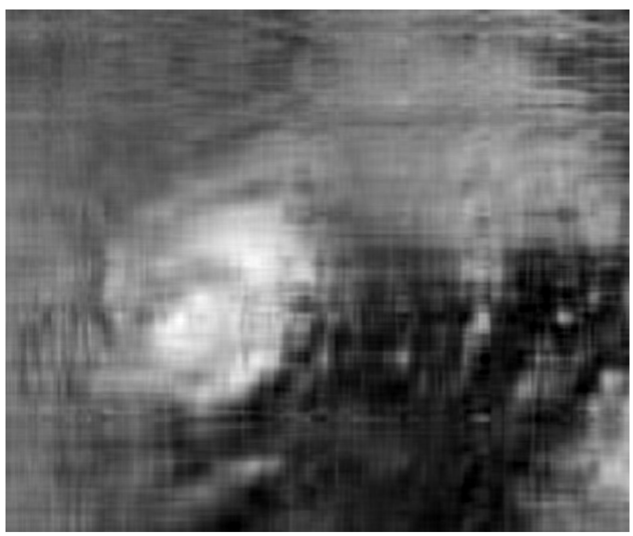

Figure 2: Approximation image for k=10

(iii) $k = 40, \frac{||X-\tilde{X}||_F}{||X||_F} = 0.083671,$

Figure 3: Approximation image for k=40

(b) The numbers to describe the approximation $\tilde{X}_k$ for $k = \{2, 10, 40\}$ are$\{5690, 28450, 113800\}$

(c) The code are shown below:

```
X = double(rgb2gray(imread('harvey-saturday-goes7am.jpg')));
[U, S, V] = svd(X);

k = [2 10 40];

for i=1:3
    app_x = zeros(size(X));
    sum_num = 0;
    for j = 1:k(i)
        app_x = app_x + S(j, j) * U(:, j) * V(:, j)';
        sum_num = sum_num + 1 + size(U(:, j)) + size(V(:, j));
    end
    s = sprintf('%d_app.jpg', k(i));
    imwrite(app_x / 256, s);
    ratio = norm(X - app_x, 'fro') / norm(X, 'fro');
    fprintf('Top %d approximation, error: %f\n', k(i), ratio);
    fprintf('numbers to store: %d\n', sum_num(1));
end
```