Sharif University of Technology
Electrical Engineering Department

Machine Learning
HW 1

Amir Hossein Yari
99102507

March 18, 2023

Contents

1. Correlation, Causality and Independence

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

$$Z \sim Uniform(a,b) \Rightarrow E[Z] = \frac{a+b}{2} \Rightarrow E[X] = \frac{-1+1}{2} = 0$$

$$E[XY] = E[X^3] = \int_{-1}^{1} x^3\, f_X(x)\, dx \overset{odd\,function}{=} 0$$

$$\Rightarrow Cov(X,Y) = 0 \Rightarrow \rho_{X,Y} = 0$$

2. Markov-Chain Gaussians

$$\rho_{X,Z} = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{Cov(X,Z)}{\sigma_X \sigma_Z}$$

$$Cov(X,Z) = E[XZ] - E[X]E[Z]$$

$$E[XZ] = E[E[XZ|Y]] = E[E[X|Y]E[Z|Y]]$$

$$E[X|Y] = \mu_X + \frac{\rho_{X,Y}\sigma_X}{\sigma_Y}(y - \mu_Y)$$

$$E[Z|Y] = \mu_Z + \frac{\rho_{Z,Y}\sigma_Z}{\sigma_Y}(y - \mu_Y)$$

$$E[XZ] = E[(\mu_X + \frac{\rho_{X,Y}\sigma_X}{\sigma_Y}(y - \mu_Y))(\mu_Z + \frac{\rho_{Z,Y}\sigma_Z}{\sigma_Y}(y - \mu_Y))]$$

$$= \mu_X\mu_Z + \frac{\rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z}{\sigma_Y^2}E[(y - \mu_Y)^2] = \mu_X\mu_Z + \rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z$$

$$Cov(X,Z) = \mu_X\mu_Z + \rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z - \mu_X\mu_Z = \rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z$$

$$\rho_{X,Z} = \frac{\rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z}{\sigma_X\sigma_Z} = \rho_{X,Y}\rho_{Z,Y}$$

3. Sensor Fusion

Mean of sensors is z because it is expected value.

$Y_1 \sim \mathcal{N}(z,\, v_1)$
$Y_2 \sim \mathcal{N}(z,\, v_2)$
$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$
$p(\mathbf{y}|z) \sim \mathcal{N}(\mathbf{y}|\mathbf{z}, \begin{bmatrix} v_1\mathbf{I} & 0 \\ 0 & v_2\mathbf{I} \end{bmatrix}) = \mathcal{N}(\mathbf{y}|Wz + b, \Sigma_{\mathbf{y}}) \Rightarrow W = \mathbf{1}\,, b = 0\,, \Sigma_{\mathbf{y}} = \begin{bmatrix} v_1\mathbf{I} & 0 \\ 0 & v_2\mathbf{I} \end{bmatrix}$
$\Rightarrow p(z|\mathbf{y}) = \mathcal{N}(z|\mu_{z|\mathbf{y}}, \Sigma_{z|\mathbf{y}})$
$\Sigma_{z|\mathbf{y}}^{-1} = \Sigma_z^{-1} + W^T\Sigma_{\mathbf{y}}^{-1}W = 0 = 0 + [\frac{1}{v_1}\mathbf{1}, \frac{1}{v_2}\mathbf{1}]\mathbf{1} = \frac{n_1}{v_1} + \frac{n_2}{v_2}$
$\Sigma_{z|\mathbf{y}} = \frac{v_1 v_2}{n_1 v_2 + n_2 v_1}$
$\mu_{z|\mathbf{y}} = \Sigma_{z|\mathbf{y}}(W^T\Sigma_{\mathbf{y}}^{-1}(y - b) + \Sigma_z^{-1}\mu_z) = \frac{v_1 v_2}{n_1 v_2 + n_2 v_1}(\frac{1}{v_1}\sum_{i=1}^{n_1}\mathbf{Y}_1^{(i)} + \frac{1}{v_2}\sum_{i=1}^{n_2}\mathbf{Y}_2^{(i)})$
$= \frac{n_1 v_2 \bar{y}_1 + n_2 v_1 \bar{y}_2}{n_1 v_2 + n_2 v_1}$

4. Pseudo Inverse

We know that $A = U\Sigma V^T = \begin{bmatrix} U_R & U_N \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix}$
$\Rightarrow A^\dagger = \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix}$

1. full row rank

So $A = \begin{bmatrix} U_R \end{bmatrix} \begin{bmatrix} S & 0 \end{bmatrix} \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix}$, $A^\dagger = \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S^{-1} \\ 0 \end{bmatrix} \begin{bmatrix} U_R^T \end{bmatrix}$
Now, we calculate $A^T(AA^T)^{-1}$ to show that this term is equal to $A^\dagger$.
$A^T(AA^T)^{-1} = \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} \begin{bmatrix} U_R \end{bmatrix} (\begin{bmatrix} U_R \end{bmatrix} \begin{bmatrix} S & 0 \end{bmatrix} \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix} \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} \begin{bmatrix} U_R \end{bmatrix})^{-1}$
$= \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} \begin{bmatrix} U_R \end{bmatrix} (U_R S^2 U_R)^{-1} = \begin{bmatrix} V_R & V_N \end{bmatrix} \begin{bmatrix} S^{-1} \\ 0 \end{bmatrix} U_R^T = A^\dagger \quad \checkmark$

2. full column rank

So $A = \begin{bmatrix} U_R & U_N \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} \begin{bmatrix} V_R^T \end{bmatrix}$, $A^\dagger = \begin{bmatrix} V_R \end{bmatrix} \begin{bmatrix} S^{-1} & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix}$

Now, we calculate $(A^T A)^{-1} A^T$ to show that this term is equal to $A^\dagger$.

$(A^T A)^{-1} A^T = (\begin{bmatrix} V_R \end{bmatrix} \begin{bmatrix} S & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix} \begin{bmatrix} U_R & U_N \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} \begin{bmatrix} V_R^T \end{bmatrix}) \begin{bmatrix} V_R \end{bmatrix} \begin{bmatrix} S & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix}$

$= (V_R S^2 V_R)^{-1} \begin{bmatrix} V_R \end{bmatrix} \begin{bmatrix} S & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix} = \begin{bmatrix} V_R \end{bmatrix} \begin{bmatrix} S^{-1} & 0 \end{bmatrix} \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix} = A^\dagger \quad \checkmark$

5. Eigenvalues

1. Tr{A} = $\sum_{i=1}^n \lambda_i$

We know that : $A \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}$

$\Rightarrow AX = \Lambda X \Rightarrow A = X \Lambda X^{-1}$

$\mathrm{Tr}\{A\} = \mathrm{Tr}\{X \Lambda X^{-1}\} \overset{trace\,is\,commutative}{=} \mathrm{Tr}\{X X^{-1} \Lambda\} = \mathrm{Tr}\{\Lambda\}$

$= \lambda_1 + \lambda_2 + ... + \lambda_n = \sum_{i=1}^n \lambda_i \quad \checkmark$

PROOF) Trace is commutative.

$\mathrm{Tr}\{AB\} = \sum_{i=1}^n \sum_{j=1}^m A_{ij} B_{ji} = \sum_{j=1}^m \sum_{i=1}^n B_{ji} A_{ij} = \mathrm{Tr}\{BA\}$

2. det{A} = $\prod_{i=1}^n \lambda_i$

$A = X \Lambda X_{-1} \Rightarrow det(A) = det(X \Lambda X_{-1}) = det(X)\, det(\Lambda)\, det(X^{-1})$

$\overset{det(X^{-1}) = \frac{1}{det(X)}}{=} det(\Lambda)$

$\lambda_1 \lambda_2 ... \lambda_n = \prod_{i=1}^n \lambda_i \quad \checkmark$

6. Maximum Likelihood Estimation

1) Bernouli

$$L(\theta) = \prod_{i=1}^{d} \theta^{x_i}\,(1-\theta)^{(1-x_i)}$$

$$Log\,L(\theta) = \sum_{i=1}^{d} Log\,\theta^{x_i}\,(1-\theta)^{(1-x_i)} = Log\,\theta\,\sum_{i=1}^{d} x_i + Log\,(1-\theta)\,\sum_{i=1}^{d}(1-x_i)$$

$$\frac{\partial Log\,L(\theta)}{\partial\theta} = 0 \Rightarrow \frac{\sum_{i=1}^{d} x_i}{\theta} - \frac{\sum_{i=1}^{d}(1-x_i)}{1-\theta} = 0 \Rightarrow \theta\,\sum_{i=1}^{d}(1-x_i) = \sum_{i=1}^{d} x_i - \theta\,\sum_{i=1}^{d} x_i$$

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{d} x_i}{d} \overset{m\,one}{=} \frac{m}{m+k}$$

2) Exponential

$$L(\lambda) = \prod_{i=1}^{d} \lambda e^{-\lambda x_i} = \lambda^d\, e^{-\lambda \sum_{i=1}^{d} x_i}$$

$$Ln\,L(\lambda) = Ln(\lambda^d) + Ln(e^{-\lambda \sum_{i=1}^{d} x_i}) = n\,Ln(\lambda) - \lambda\,\sum_{i=1}^{d} x_i$$

$$\frac{\partial Ln\,L(\lambda)}{\partial\lambda} = 0 \Rightarrow \frac{n}{\lambda} - \sum_{i=1}^{d} x_i = 0$$

$$\hat{\lambda}_{MLE} = \frac{d}{\sum_{i=1}^{d} x_i}$$

3) Normal

$$L(\mu,\sigma^2) = \prod_{i=1}^{d} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{d}(x_i-\mu)^2}$$

$$Ln\, L(\mu,\sigma^2) = Ln((2\pi\sigma^2)^{-\frac{n}{2}}) + Ln(e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{d}(x_i-\mu)^2}) = -\frac{n}{2}\, Ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{d}(x_i-\mu)^2$$

$$\frac{\partial Ln\, L(\mu,\sigma^2)}{\partial\mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{d}(x_i-\mu) = 0 \Rightarrow \sum_{i=1}^{d} x_i - n\mu = 0$$

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{d} x_i}{d}$$

$$\frac{\partial Ln\, L(\mu,\sigma^2)}{\partial\sigma^2} = 0 \Rightarrow -\frac{n}{2\sigma^2} - (\frac{1}{2}\sum_{i=1}^{d}(x_i-\mu)^2)\frac{-1}{\sigma^4} = \frac{1}{2\sigma^2}(\frac{1}{\sigma^2}\sum_{i=1}^{d}(x_i-\mu)^2 - n) = 0$$

$$\hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^{d}(x_i-\mu)^2}{d}$$

7. A Tiny Bit of Vector Differentiation

1. $\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T$ **or a**

$$\mathbf{a}^T\mathbf{x} = [a_1 a_2 \ldots a_d]\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = a_1x_1 + a_2x_2 + \cdots + a_dx_d$$

$$\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = [\tfrac{\partial\mathbf{a}^T\mathbf{x}}{\partial x_1}\ \tfrac{\partial\mathbf{a}^T\mathbf{x}}{\partial x_1}\ \ldots\ \tfrac{\partial\mathbf{a}^T\mathbf{x}}{\partial x_d}]^T = [a_1\, a_2\, \ldots\, a_d]^T = \mathbf{a} \qquad \checkmark$$

2. $\nabla_{\mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = \mathbf{x}^T(A + A^T) = (A + A^T)\mathbf{x}$

$$\mathbf{x}^T A\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \ldots & x_d \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1d} \\ a_{21} & a_{22} & \ldots & a_{2d} \\ \vdots & \vdots & & \vdots \\ a_{d1} & a_{d2} & \ldots & a_{dd} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}x_ix_j$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = \begin{bmatrix} \frac{\partial\mathbf{x}^T A\mathbf{x}}{\partial x_1} & \frac{\partial\mathbf{x}^T A\mathbf{x}}{\partial x_2} & \ldots & \frac{\partial\mathbf{x}^T A\mathbf{x}}{\partial x_d} \end{bmatrix}^T$$

$\frac{\partial \mathbf{x}^T A\mathbf{x}}{\partial x_k} = \frac{\partial}{\partial x_k}(x_1 \sum_{i=1}^{d} a_{i1}x_i + \cdots + x_k \sum_{i=1}^{d} a_{ik}x_i + x_d \sum_{i=1}^{d} a_{id}x_i)$
$= x_1 a_{k1} + \cdots + (\sum_{i=1}^{d} a_{ik}x_i + a_{kk}x_k) + \cdots + x_d a_{kd}$
$= \sum_{j=1}^{d} a_{kj}x_j + \sum_{i=1}^{d} a_{ik}x_i (row\, k\, A \times \mathbf{x} + row\, k\, A^T \times \mathbf{x}) = A\mathbf{x} + A^T\mathbf{x} = (A + A^T)\mathbf{x}\checkmark$

8. Bayes Rule for Gaussian Variables

First, we define the X_1' variable as follows.

$X_1' = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$

X_1', X_2 are jointly MVN.

$Cov(X_1', X_2) = Cov(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2, X_2) = Cov(X_1, X_2) - \Sigma_{12}\Sigma_{22}^{-1}Cov(X_2, X_2) = 0$

So X_1, X_2 are uncorrelated. So they are Independent.

$E[X_1'|X_2 = x_2] \overset{Independent}{=} E[X_1'] = E[X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2] = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$
$E[X_1|X_2 = x_2] = E[X_1'|X_2 = x_2] + \Sigma_{12}\Sigma_{22}^{-1}x_2 = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2$
$\Rightarrow \mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$

We know that X_1', X_2 are independent.

$\Rightarrow Cov(X_1'|X_2 = x_2) = Cov(X_1')$
$Cov(X_1'|X_2 = x_2) = Cov(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2|X_2 = x_2) = Cov(X_1|X_2 = x_2)$ (1)
$Cov(X_1') = Cov(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2) \overset{\star}{=} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (2)
$\xrightarrow{(1)\,\&\,(2)} Cov(X_1|X_2 = x_2) = \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

The reason for equality($\star$) is that we have in general:

$Cov(X_2 - DX_1) = Cov(X_2, X_2) - D\,Cov(X_1, X_2) - Cov(X_2, X_1)\,D^T + D\,Cov(X_1, X_1)\,D^T$

9. Implicit Regularization!

1. $\mathbf{x}^\star = \mathbf{A}^\dagger\mathbf{b}$

Considering that the number of variables is more than the number of equations ($n > m$), it can be said simply there are multiple answers for the problem.

Goal : minimize $||\mathbf{x}||^2$ subject to $\mathbf{Ax} = \mathbf{b}$
Suppose $\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{z}$ and $\mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$. For any $\mathbf{x}$ satisfies $\mathbf{Ax} = \mathbf{b}$, we have:
$||\mathbf{x}||^2 = ||\mathbf{x}-\hat{\mathbf{x}}+\hat{\mathbf{x}}||^2 = ||\mathbf{x}-\hat{\mathbf{x}}||^2+||\hat{\mathbf{x}}||^2+2\,\hat{\mathbf{x}}^T(\mathbf{x}-\hat{\mathbf{x}}) = ||\mathbf{x}-\hat{\mathbf{x}}||^2+||\hat{\mathbf{x}}||^2+2\,\mathbf{z}^T\mathbf{A}(\mathbf{x}-\hat{\mathbf{x}})$
$= ||\mathbf{x}-\hat{\mathbf{x}}||^2 + ||\hat{\mathbf{x}}||^2 \geq ||\hat{\mathbf{x}}||^2$
$\mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{z} \Rightarrow \mathbf{A}\mathbf{A}^T\mathbf{z} = \mathbf{b}$
So if $\mathbf{A}\mathbf{A}^T$ is invertible($\mathbf{A}$ has linearly independent rows) $\Rightarrow \hat{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}b$
$= \mathbf{A}^\dagger\mathbf{b} \quad \checkmark$

2. gradient descent method

if $F(\mathbf{x}) = ||\mathbf{Ax}-\mathbf{b}||^2 = (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\ \mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}\ \mathbf{x} - \mathbf{b}^T\mathbf{b}$
$\nabla_\mathbf{x}F(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A}\ \mathbf{x} - 2\mathbf{A}^T\mathbf{b} \Rightarrow \nabla_\mathbf{x}F(\mathbf{x}^{(t)}) = 2\mathbf{A}^T(\mathbf{Ax}^{(t)} - \mathbf{b})$
$\Rightarrow \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \epsilon\mathbf{A}^T(\mathbf{Ax}^{(t)} - \mathbf{b}) = \mathbf{x}^{(t)} - \frac{\epsilon}{2}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})$
We know that the gradient is in the direction of the greatest increase of a function. So, the negative of the gradient is in the direction of the greatest decrease of the function. Given that $||\mathbf{Ax}-\mathbf{b}||^2 \geq 0$, we have:
$F(\mathbf{x}^{(0)}) \geq F(\mathbf{x}^{(1)}) \geq F(\mathbf{x}^{(2)}) \geq \cdots \geq 0$ (with appropriate value for ϵ)
Considering the convexity of the function($F(\mathbf{x})$) and above sequence, it can be said that $F(\mathbf{x}^{(t)})$ will converge to the $\min\limits_{\mathbf{x}^{(t)}} F(\mathbf{x}^{(t)})$.
So $\mathbf{x}^{(t)}$ will converge to $\mathbf{x}^\star$. $\checkmark$

3. upper bound of learning rate

We know that $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\epsilon}{2}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})$
$||\mathbf{Ax}^{(t+1)} - \mathbf{b}||^2 = ||\mathbf{Ax}^{(t)} - \frac{\epsilon}{2}\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)}) - \mathbf{b}||^2$
$= (\mathbf{Ax}^{(t)} - \frac{\epsilon}{2}\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)}) - \mathbf{b})^T(\mathbf{Ax}^{(t)} - \frac{\epsilon}{2}\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)}) - \mathbf{b})$
$= (\mathbf{Ax}^{(t)}-\mathbf{b})^T(\mathbf{Ax}^{(t)}-\mathbf{b}) - \frac{\epsilon}{2}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T(\mathbf{Ax}^{(t)}-\mathbf{b}) - \frac{\epsilon}{2}(\mathbf{Ax}^{(t)}-\mathbf{b})^T\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)}) +$
$\frac{\epsilon^2}{4}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})$
$= ||\mathbf{Ax}^{(t)} - \mathbf{b}||^2 - \frac{\epsilon}{2}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T(\mathbf{Ax}^{(t)} - \mathbf{b}) - \frac{\epsilon}{2}(\mathbf{Ax}^{(t)} - \mathbf{b})^T\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})$
$+ \frac{\epsilon^2}{4}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})$

Now we should prove:
$\frac{\epsilon^2}{4}\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T\mathbf{A}\nabla_\mathbf{x}F(\mathbf{x}^{(t)}) \leq \epsilon\nabla_\mathbf{x}F(\mathbf{x}^{(t)})^T\mathbf{A}^T(\mathbf{Ax}^{(t)} - \mathbf{b})$

Assuming that ϵ is positive, we have:

$\nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T \mathbf{A} \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)}) \leq \frac{4}{\epsilon} \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T (\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b})$

By assuming $\epsilon \leq \frac{2}{\sigma_{max}^2(A)}$:
$\frac{4}{\epsilon} \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T (\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}) \geq 2\sigma_{max}^2(A) \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T (\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b})$
$= \sigma_{max}^2(A) \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})$

Now we should prove:
$\nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T \mathbf{A} \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)}) \leq \sigma_{max}^2(A) \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})$
$\Rightarrow \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \mathbf{A}^T \mathbf{A} \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)}) - \sigma_{max}^2(A) \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)}) \leq 0$
$\Rightarrow \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})^T (\sigma_{max}^2(A)\mathbf{I} - \mathbf{A}^T \mathbf{A}) \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)}) \geq 0$

So if $\sigma_{max}^2(A)\mathbf{I} - \mathbf{A}^T\mathbf{A}$ be positive semi-definite, the verdict will be confirmed.
$\mathbf{A} = U\Sigma V^T \Rightarrow \mathbf{A}^T\mathbf{A} = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T \Sigma V^T = V\Sigma^2 V^T$

$$\Sigma^2 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & \sigma_r^2 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & 0 & \dots & 0 \end{bmatrix} \qquad \Sigma_1^2 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_1^2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & \sigma_1^2 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & 0 & \dots & \sigma_1^2 \end{bmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \dots \sigma_r$

$\Rightarrow \sigma_{max}^2(A)\mathbf{I} - \mathbf{A}^T\mathbf{A} = V\Sigma_1^2 V^T - V\Sigma^2 V^T = V(\Sigma_1^2 - \Sigma^2)V^T$

$$= V \begin{bmatrix} 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_1^2 - \sigma_2^2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & \sigma_1^2 - \sigma_r^2 & \dots & 0 \\ \vdots & \vdots & & & & \\ 0 & 0 & \dots & 0 & \dots & \sigma_1^2 \end{bmatrix} V^T$$

So eigenvalues of $(\sigma_{max}^2(A)\mathbf{I} - \mathbf{A}^T\mathbf{A})$ are $\{0, \sigma_1^2 - \sigma_2^2, \dots, \sigma_1^2 - \sigma_r^2, \sigma_1^2, \dots, \sigma_1^2\}$ that all of them are greater than or equal zero. It means that $(\sigma_{max}^2(A)\mathbf{I} - \mathbf{A}^T\mathbf{A})$ is positive semi-definite. $\checkmark$