EECS 553 Homework 2 Solution (FA24)

1. **Full Rank Matrices** (3 points)

   Grading Rubrics

   1. 3 points for fully correct answer with proper justification. Partial credit – Add one point for each:
      - 1 point for performing SVD on $\boldsymbol{B}$.
      - 1 point for showing $\Longrightarrow$.
      - 1 point for showing $\Longleftarrow$.
   2. 0 if no effort.

   Let $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a thin SVD of $\boldsymbol{B}$ where $\boldsymbol{U} \in \mathbb{R}^{p\times q}$, $\boldsymbol{V} \in \mathbb{R}^{q\times q}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$ and is diagonal. So $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T$ is also an SVD.

   $\text{rank}(\boldsymbol{B}) = q \iff$ all elements on the diagonal of $\boldsymbol{\Sigma}$ are nonzero $\iff$ all elements on the diagonal of $\boldsymbol{\Sigma}^2$ are nonzero $\iff \boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T$ is invertible.

   Note: Alternative solution may receive full or partial credit.

   Alternative Solution: B has full rank $\iff \forall \mathbf{x} \neq \mathbf{0}, B\mathbf{x} \neq \mathbf{0} \iff \forall \mathbf{x} \neq \mathbf{0}, \|B\mathbf{x}\| > 0 \iff \forall \mathbf{x} \neq \mathbf{0}, \|B\mathbf{x}\|^2 > 0 \iff \forall \mathbf{x} \neq \mathbf{0}, \mathbf{x}^T B^T B\mathbf{x} > 0 \iff B^T B$ is PD $\iff B^T B$ is invertible (since $B^T B$ is symmetric),

2. **Weighted Least Squares** (3 points)

   Grading Rubrics

   1. 3 points for fully correct answer with proper justification. Partial credit – Add one point for each:
      - 1 point for reformulating the problem as a quadratic programming problem.
      - 1 point for identifying the convexity of the problem.
   2. 0 if no effort

   Let

   $$\boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^T \\ 1 & \boldsymbol{x}_2^T \\ \dots & \\ 1 & \boldsymbol{x}_n^T \end{bmatrix} \tag{1}$$

   and

   $$\boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} \tag{2}$$

   The objective can be written as $\boldsymbol{J}(\boldsymbol{\theta}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^T\boldsymbol{C}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{\theta}^T(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}$.

$\nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\theta}) = 2(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})\boldsymbol{\theta} - 2\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{y}$.

$\nabla^2_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\theta}) = 2(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})$.

$\nabla^2_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\theta})$ is PSD, so the objective is convex. The solution can be determined letting $\nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\theta}) = 0$, leading to $\boldsymbol{\theta}^* = (\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{y})$, assuming $\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X}$ is invertible.

3. Grading Rubrics

1. 3 points for fully correct answer with proper justification. Partial credit – Add one point for each:
   - 1 point for correct statements on the probabilistic assumptions.
   - 1 point for applying the Bayes rule.
   - 1 point for correct w and b.
2. 0 if no effort.

The LDA model assumes a normal distribution for the class-conditional densities $f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y = k)$, for $k = 0, 1$. On the other hand, logistic regression assumes

$$\eta(\boldsymbol{x}) = \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp\{-(\boldsymbol{w}^T\boldsymbol{x} + b)\}}$$

Now, we show that the distributional assumption of LDA implies the distributional assumption of logistic regression in the case of binary classification. Using Bayes rule, we have

$$\begin{aligned}
\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) &= \frac{f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y = 1)\Pr(Y = 1)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \frac{f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y = 1)\Pr(Y = 1)}{f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y = 1)\Pr(Y = 1) + f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y = 0)\Pr(Y = 0)} \\
&= \frac{1}{1 + \frac{f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y=0)\Pr(Y=0)}{f_{\boldsymbol{X}|Y}(\boldsymbol{x}|Y=1)\Pr(Y=1)}}
\end{aligned}$$

Let us denote by π_0, π_1 the *a priori* probabilities of the class label Y. Then, using the LDA assumption that $f_{X|Y}(\boldsymbol{x}|Y = k)$ for $k \in \{0, 1\}$ is normally distributed according to $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, we get

$$\begin{aligned}
\Pr(Y = 1|X = \boldsymbol{x}) &= \frac{1}{1 + \frac{\pi_0 \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}}{\pi_1 \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\}}} \\
&= \frac{1}{1 + \frac{\pi_0}{\pi_1}\exp\{-[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \frac{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{2}]\}} \\
&= \frac{1}{1 + \exp\{-[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \frac{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{2} - \log\frac{\pi_0}{\pi_1}]\}}
\end{aligned}$$

Thus showing that $\Pr(Y = 1|X = \boldsymbol{x})$ is of the form $\frac{1}{1+\exp\{-(\boldsymbol{w}^T\boldsymbol{x}+b)\}}$, where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and $b = \frac{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{2} - \log\frac{\pi_0}{\pi_1}$.

4. Logistic regression objective function (3 pts each)

(a) **Grading Rubrics**

1. 3 points for fully correct answers/reasoning. Some partial credits you can (but not required to, as long as the final answer is essentially correct) consider:
   - 1 point for showing correct $P(y|\tilde{\boldsymbol{x}};\boldsymbol{\theta})$
   - 0.5 point each for showing the correct probability for $y=-1$ and $y=1$
   - 1 point for showing the correct loss function
2. 0 if no effort

Recall logistic regression is assuming the following likelihood function:

$$\begin{aligned} P(y=1|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) &= \frac{1}{1+e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \\ P(y=-1|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) &= \frac{e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}}{1+e^{-\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \\ &= \frac{1}{1+e^{\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}} \end{aligned}$$

Alternatively we can write:

$$P(y|\tilde{\boldsymbol{x}};\boldsymbol{\theta}) = \frac{1}{1+e^{-y\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}}}$$

Thus the negative log-likelihood function:

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\log P(y_i|\tilde{\boldsymbol{x}}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n}\log(1+\exp(-y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i)$$

Hence with the new notation of $\phi(t)=\log(1+\exp(-t))$, the logistic regression regularized negative log-likelihood may be written

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n}\phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) + \lambda\|\boldsymbol{w}\|^2.$$

(b) Grading Rubrics

1. 3 points for correct gradient (in either form given below)
2. 2 points for answers that only make minor mistakes including: sign errors, missing regularization terms, etc.
3. 1 point if mostly wrong
4. 0 if no effort

First by chain rule, we have:

$$\nabla_{\boldsymbol{\theta}}\phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) = \phi'(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i)y_i\tilde{\boldsymbol{x}}_i$$

where $\phi'(t) = \frac{-\exp(-t)}{1+\exp(-t)} = -\frac{1}{1+\exp(t)}$
Then by linearity of gradient, we have:

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= 2\lambda[0,\boldsymbol{w}^T]^T + \sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\phi(y_i\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i) \\ &= 2\lambda[0,\boldsymbol{w}^T]^T - \sum_{i=1}^{n}y_i(\frac{1}{1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)})\tilde{\boldsymbol{x}}_i \end{aligned}$$

Alternatively answer:

$$\nabla J(\boldsymbol{\theta}) = 2\lambda[0, \boldsymbol{w}^T]^T - \sum_{i=1}^{n} y_i \big(\frac{\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}{1+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}\big)\tilde{\boldsymbol{x}}_i$$

(c) **Grading Rubrics**

1. 3 points for correct Hessian
2. 2 points for answers that only make minor mistakes including: sign errors, missing regularization terms, etc.
3. 1 point if mostly wrong
4. 0 point if no effort

The Hessian

$$\begin{aligned}
\mathbf{H} &= \frac{\partial}{\partial\boldsymbol{\theta}^T}\left(\frac{\partial J(\theta)}{\partial\boldsymbol{\theta}}\right) \\
&= \frac{\partial}{\partial\theta^T}\left\{2\lambda[0, \boldsymbol{w}^T]^T - \sum_{i=1}^{n} y_i \big(\frac{1}{1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}\big)\tilde{\boldsymbol{x}}_i\right\} \\
&= 2\lambda\begin{bmatrix}0 & 0\\ 0 & \mathcal{I}\end{bmatrix} + \sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i{}^T y_i^2\left(\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\left[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]^2}\right) \\
&= 2\lambda\begin{bmatrix}0 & 0\\ 0 & \mathcal{I}\end{bmatrix} + \sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i{}^T\left(\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\left[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]^2}\right)
\end{aligned}$$

where $\mathcal{I}$ is a $d \times d$ identity matrix.

Note:

$$\begin{aligned}
\frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\left[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]^2} &= \frac{1}{\left[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]\left[1+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]} \\
&= \frac{1}{2+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)+\exp(-y_i\theta^T\tilde{\boldsymbol{x}}_i)}
\end{aligned}$$

So any form of above are correct answers.

(d) **Grading Rubrics**

1. 3 points for fully correct answer that associated the (semi) positive-definiteness of the Hessian to λ. Some partial credits you can (but not required to, as long as the final answer is essentially correct) consider:
   - 1 point for computing bilinear form $\boldsymbol{z}^T\mathbf{H}\boldsymbol{z}$ correctly
   - 1 point for each for correctly concluding convex cases and strictly convex cases respectively
2. 0 point if no effort

Letting $a_i = \frac{\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)}{\left[1+\exp(y_i\theta^T\tilde{\boldsymbol{x}}_i)\right]^2} > 0$ regardless of $\tilde{\boldsymbol{x}}_i$ and y_i, we have for any $\boldsymbol{z} \in \mathbb{R}^{d+1}$ such that $\boldsymbol{z} \neq 0$:

$$\begin{aligned}
\boldsymbol{z}^T \mathbf{H} \boldsymbol{z} &= \boldsymbol{z}^T \left(\sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^T a_i + 2\lambda \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{I} \end{bmatrix} \right) \boldsymbol{z} \\
&= \sum_{i=1}^{n} a_i (\boldsymbol{z}^T \tilde{\boldsymbol{x}}_i)(\tilde{\boldsymbol{x}}_i^T \boldsymbol{z}) + 2\lambda(\boldsymbol{z}^T \boldsymbol{z} - z_1^2) \\
&= \textstyle\sum_{i=1}^{n} a_i (\boldsymbol{z}^T \tilde{\boldsymbol{x}}_i)^2 + 2\lambda(\|\boldsymbol{z}\|^2 - z_1^2)
\end{aligned}$$

Observe:
1) when $\lambda \geq 0$, we have $\boldsymbol{z}^T \mathbf{H} \boldsymbol{z} \geq 0, \forall \boldsymbol{z}$ (i.e Hessian is PSD everywhere), hence the problem is convex.
2) when $\lambda > 0$, $\forall \boldsymbol{z} \neq 0$, if for all i, $z_i = 0$ excpet for $z_1 \neq 0$, $\boldsymbol{z}^T \mathbf{H} \boldsymbol{z} > \sum_{i=1}^{n} a_i (\boldsymbol{z}^T \tilde{\boldsymbol{x}}_i)^2 = z_1^2 \sum_{i=1}^{n} a_i > 0$. Otherwise, $\boldsymbol{z}^T \mathbf{H} \boldsymbol{z} \geq 2\lambda(\|\boldsymbol{z}\|^2 - z_1^2) = 2\lambda\|\boldsymbol{z}\|^2 > 0$.

5. Logistic Regression for Fashion Classification (3 points each)

(a) **Grading Rubrics**

1. 3 points if fully correct:
   - 1 point for correct test error (within $\pm 1\%$)
   - 1 point for correct number of iterations (within ± 1)
   - 1 point for correct reported objective function after convergence (within ± 10)
2. 0 point if no effort

Test error = 3.6%
Number of iterations = 8
Value of objective function after convergence = 451.2632670172336
hw2 Note: the values might not exactly match. Answers within 1% of test error, ± 1 number of iterations, or ± 10 for the objective are acceptable.

(b) **Grading Rubrics**

1. 3 points if fully correct:
   - 1 point for explaining what is the notion of confidence used (anything related to distance from boundary or value of $x \cdot \theta + b$)
   - 2 points for showing the 20 misclassified examples (figures may not be exactly the same, give full credits for submissions that resemble the solution)
2. 0 point if no effort

See Figure 1 for the figure of the misclassified images. We define confidence as the distance to the learned hyperplane. The further a point is away from the hyperplane, the more confident the classifier is.

(c) **Grading Rubrics**

1. 3 points for submitting code (as long as students are not submitting only print functions, full credits will be given)
2. 0 point if not code submitted

Figure 1: P5 Figure