1 Problem 1

1.1 Part a

1.1.1 i: Given an invertible matrix A, $(A^{-1})^T = (A^T)^{-1}$

Given that a full-rank matrix A is invertible, we can deduce two important facts:

- A^T is invertible
- The inverses for both A and A^T are unique

Therefore, my proof is as follows. If I can show both $(A^{-1})^T$ and $(A^T)^{-1}$, are valid inverses for A^T, then I can conclude $(A^{-1})^T = (A^T)^{-1}$ via the above two facts.

The proof for $(A^T)^{-1}$ is trivial, since it is, by definition, the inverse of A^T. Since A^T is full rank, it will have an inverse and

$$A^T(A^T)^{-1} = I \tag{1}$$

Now, I aim to show that $A^T(A^{-1})^T = I$. By the properties of transposes, we see

$$A^T(A^{-1})^T = (AA^{-1})^T = I \tag{2}$$

Since A is full rank, $AA^{-1} = I$ and the transpose of the identity is the identity. Therefore, since both $(A^{-1})^T$ and $(A^T)^{-1}$, are valid inverses for A^T, it can be concluded $(A^{-1})^T = (A^T)^{-1}$ and the proof is complete.

1.1.2 ii: Given that matrix A, B, $A + B$ are invertible, $(A + B)^{-1} = A^{-1} + B^{-1}$

To disprove this claim, I will first show a symbolic proof, then a counterexample. First off, it's pretty easy to see that since $A + B$ is invertible, its inverse will be unique. Next, it's also trivial to see

$$(A + B)(A + B)^{-1} = I \tag{3}$$

Therefore, in order for the above statement to hold it must be true that $(A + B)(A^{-1} + B^{-1}) = I$. However,

$$\begin{aligned}(A + B)(A^{-1} + B^{-1}) &= AA^{-1} + AB^{-1} + BA^{-1} + BB^{-1} &(4)\\ &= 2I + AB^{-1} + BA^{-1} \neq I \;\; \forall A, B \; invertible &(5)\end{aligned}$$

As a counterexample, consider the case when $A, B = I$. Equation (4) will be, since $I^{-1} = I$,

$$(A + B)(A^{-1} + B^{-1}) = (2I)(2I) = 4I \neq I \tag{6}$$

1.1.3 iii: The inverse of a symmetric matrix is itself symmetric

So we are asked to prove that if $A = A^T$, $A^{-1} = (A^{-1})^T$. Therefore, to prove this claim it is sufficient to show $AA^{-1} = A(A^{-1})^T = I$. This is sufficient since we are given that the matrix A is symmetric, so $A = A^T$ and the above two expressions aim to show $A^TA^{-1} = A^T(A^{-1})^T = I$.

Therefore, our goal becomes to prove $AA^{-1} = A(A^{-1})^T = I$. We can then separate this into two categories

A is not invertible If A is not invertible, then A^{-1} does not exist and this fact becomes trivially false. Since no inverse exists, this fact cannot be proved true or false.

A is invertible If A is invertible, the proof is significantly less trivial. It is easy to see that $AA^{-1} = I$, and now I must show $A(A^{-1})^T = I$. To do so, see

$$\begin{aligned} A(A^{-1})^T &= A^T(A^{-1})^T && \text{since } A = A^T && (7)\\ &= (AA^{-1})^T && \text{due to properties of transpose} && (8)\\ &= I^T = I && && (9) \end{aligned}$$

and the proof is complete. Since the inverse of A is unique, we can say $A^{-1} = (A^{-1})^T$.

Therefore, this statement can be proved given that A is full rank. Technically, a matrix full of zeros is invertible, but obviously this matrix would not have an inverse.

1.2 Part b

In this problem, we are given a $m \times n$ matrix X with can be decomposed, via singular values, as $X = U\Sigma V^T$. We also know $UU^T = U^TU = VV^T = V^TV = I$ and Σ contains non-increasing non-negative values along its diagonal and zeros elsewhere. We are then asked to compute the eigendecomposition of XX^T.

To do so, see

$$\begin{aligned} XX^T &= U\Sigma V^T \left(U\Sigma V^T\right)^T && \text{By SVD definition} && (10)\\ &= U\Sigma V^T V \Sigma^T U^T && && (11)\\ &= U\Sigma\Sigma^T U^T && \text{By definition of } V && (12) \end{aligned}$$

Therefore, we define $\Lambda = \Sigma\Sigma^T$ and $Q = U$. Since U is square and $UU^T = I$, we can define $Q^{-1} = U^T$. So the eigenvalues of XX^T lie in the diagonal matrix Λ and the eigenvectors lie in the columns of the square matrix U.

1.3 Part c

This problem was done in Python. Since the code is not required, I will merely show the results for both part a and b. The screenshot for the Python code is shown below where both answers are shown:

```
[Anaconda2] C:\Users\erwarner\Dropbox\Winter 2016\EECS 545\Homework 1>hw1.py
The first 3 singular values are: [ 757.00219118  158.20556896  130.31529335]
The Frobenius norm of A-B is: 261.00882134
```

[scale=1]

2 Problem 2

2.1 Part a

2.1.1 i: $P(H = h|D = d)$? $P(H = h)$

From Bayes, we know

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

substituting $H = h$ for A and $D = d$ for B, we see

$$P(H = h|D = d) = \frac{P(D = d|H = h)}{P(D = d)} P(H = h) \tag{14}$$

And therefore it is easy to see that the relationship between $P(H = h|D = d)$ and $P(H = h)$ is dependent on the ratio $\frac{P(D=d|H=h)}{P(D=d)}$. If this ratio is less than or equal to one, we can conclude $P(H = h|D = d) \leq P(H = h)$ with analogous relationships holding for $\geq$ and $=$. However, there is no way to determine the value of this ratio. It could very well be the case that $P(D = d)$ is very low, but when we know $H = h$, $P(D = d|H = h)$ gets much larger. In this case, the sign would be $\geq$. It could also be the case that $P(D = d)$ is very high, but when we know $H = h$, $P(D = d|H = h)$ gets much lower. In this case, the sign would be $\leq$. Therefore, we can only say that the question mark in $P(H = h|D = d)$? $P(H = h)$ is **depends**.

2.1.2 ii: $P(H = h|D = d)$? $P(D = d|H = h)P(H = h)$

Once again, we go back to Bayes Theorem, which says

$$P(H = h|D = d) = \frac{P(D = d|H = h)P(H = h)}{P(D = d)} \tag{15}$$

Now, the question mark only depends on $\frac{1}{P(D=d)}$. From the axioms of probability, we know that $P(D = d) \leq 1$ This means that, due to this fact,

$$P(H = h|D = d) \leq P(D = d|H = h)P(H = h) \tag{16}$$

2.2 Part b

In this problem we are given random variables X and Y which have a joint distribution $p(x, y)$.

2.2.1 i: $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$

From definition, we know

$$\mathbb{E}_X[X|Y]] = \int \frac{xp(x, y)}{p_y(y)} dx \tag{17}$$

Therefore, $\mathbb{E}_Y$ of the above would be:

$$\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \int \left(\int \frac{xp(x, y)}{p_y(y)} dx \right) p_y(y) dy \tag{18}$$

To start, since $p_y(y)$ is non-dependent on x, we can pull that term out of the dx term. This means

$$= \int \left(\int xp(x,y)dx \right) \frac{p_y(y)}{p_y(y)} dy = \int \left(\int xp(x,y)dx \right) dy \tag{19}$$

Lastly, due to the properties of integrals, we know:

$$\int \left(\int xp(x,y)dx \right) dy = \int \left(\int p(x,y)dy \right) xdx \tag{20}$$

And we know $\int p(x,y)dy = p_x(x)$, so therefore the integral becomes:

$$\int \left(\int p(x,y)dy \right) xdx = \int p_x(x)xdx \tag{21}$$

which, by definition, $\mathbb{E}[X] = \int p_x(x)xdx$ and the result is proved.

2.2.2 ii: $\mathbf{var}[X] = \mathbb{E}_Y[\mathbf{var}_X[X|Y]] + \mathbf{var}_Y[\mathbb{E}_X[X|Y]]$

One important fact from class is that:

$$var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \tag{22}$$

With this fact, we can expand $\mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]]$ to

$$\mathbb{E}_y\left(\mathbb{E}_x[([X|Y] - \mathbb{E}_x[X|Y])^2]\right) + \mathbb{E}_y\left(\{\mathbb{E}_x[X|Y] - \mathbb{E}_y\mathbb{E}_x[X|Y]\}^2\right) \tag{23}$$

From part i of this problem, we know $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$ and therefore the below can be reformulated as

$$\mathbb{E}_y\left(\mathbb{E}_x[([X|Y] - \mathbb{E}_x[X|Y])^2]\right) + \mathbb{E}_y\left([\mathbb{E}_x[X|Y] - \mathbb{E}[X]]^2\right) \tag{24}$$

$$= \mathbb{E}_y\left\{\mathbb{E}_x([X|Y]^2) - 2\mathbb{E}_x[X|Y]\mathbb{E}_x[X|Y] + (\mathbb{E}_x[X|Y])^2 + (\mathbb{E}_x[X|Y])^2 - 2\mathbb{E}_x[X|Y]\mathbb{E}[X] + \mathbb{E}[X]^2\right\} \tag{25}$$

$$= \mathbb{E}_y\left\{\mathbb{E}_x([X|Y]^2) - 2\mathbb{E}_x[X|Y]\mathbb{E}[X] + \mathbb{E}[X]^2\right\} \quad \text{Move the } \mathbb{E}_y \text{ inside} \tag{26}$$

$$= \mathbb{E}_y\mathbb{E}_x([X|Y]^2) - 2\mathbb{E}_y\mathbb{E}_x[X|Y]\mathbb{E}[X] + \mathbb{E}[X]^2 \quad \text{since } \mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]] \tag{27}$$

$$= \mathbb{E}_y\mathbb{E}_x([X|Y]^2) - (\mathbb{E}_x[X])^2 \tag{28}$$

So what we must prove is $\mathbb{E}_y\mathbb{E}_x([X|Y]^2) = \mathbb{E}(X^2)$. From class, we know this term is:

$$\int \left(\int x^2 p(x|y)dx \right) p_y(y)dy = \int \left(\int x^2 \frac{p(x,y)}{p_y(y)} dx \right) p_y(y)dy \tag{29}$$

Using the logic from part i, this integral can be reduced to:

$$\int x^2 p_x(x)dx = \mathbb{E}(X^2) \tag{30}$$

which means (28) is

$$\mathbb{E}(X^2) - (\mathbb{E}[X])^2 = var[X] \tag{31}$$

and the theorem is proved.

3 Problem 3

Using the spectral decomposition, we know that $Au_i = \lambda_i u_i$, where u_i is the i^{th} column of a $d \times d$ matrix U such that $U^T U = I$ and $A = U\Lambda U^T$. λ_i is the i^{th} eigenvalue and $\Lambda = diag(\lambda_1, \ldots, \lambda_d)$. Since $U^T U = I$, we can say

$$u_j^T u_i = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{32}$$

3.1 Part a: Prove A is PSD $\Leftrightarrow \lambda_i \geq 0$ for each i

3.1.1 $\Rightarrow$

For now, we assume that A is PSD, which means for all $x \in \mathbb{R}^d$, $x^T Ax \geq 0$. Additionally, we know that

$$Au_i = \lambda_i u_i \tag{33}$$

For each eigenvalue $diag(\lambda_1, \ldots, \lambda_d)$. Multiplying each side by u_i^T we get

$$u_i^T Au_i = u_i^T \lambda_i u_i \qquad \text{since } \lambda_i \text{ is a scalar} \tag{34}$$

$$u_i^T Au_i = u_i^T u_i \lambda_i = \lambda_i \tag{35}$$

So we get the important relation $u_i^T Au_i = \lambda_i$, and we know, from the assumption A is PSD, that $u_i^T Au_i \geq 0$ for any u_i. Therefore, $0 \leq u_i^T Au_i = \lambda_i$ and $\lambda_i \geq 0$ in order for the relation $Au_i = \lambda_i u_i$ to hold. This will hold for any $i \in \{1, \ldots, d\}$, and therefore we can say $\lambda_i \geq 0$ for $i \in \{1, \ldots, d\}$ and the theorem is proved.

3.1.2 $\Leftarrow$

Now, we assume that $\lambda_i \geq 0$ for each i, and must prove that $x^T Ax \geq 0$ for all $x \neq 0$. To complete this proof, take any vector x and define it as $x = U\tilde{x}$. Since $UU^T = I$, we know U has an inverse and therefore is full rank. Therefore, $x = U\tilde{x}$ will not suffer any loss of rank and can span $\mathbb{R}^{d\times 1}$. Using this, it is easy to see:

$$x^T Ax = \tilde{x}^T U^T AU\tilde{x} \qquad \text{from eigendecomposition know } U^T AU = \Lambda \tag{36}$$

$$= \tilde{x}^T \Lambda \tilde{x} \tag{37}$$

Since we know $\Lambda \geq 0$, by definition of a positive semi-definite matrix $\tilde{x}^T \Lambda \tilde{x} \geq 0$ and therefore

$$x^T Ax = \tilde{x}^T \Lambda \tilde{x} \geq 0 \tag{38}$$

and therefore, via the equality $x^T Ax \geq 0$ and the proof is complete.

3.2 Part b: Prove A is PD $\Leftrightarrow \lambda_i > 0$ for each i

The proof for both necessity and sufficiency follow the exact same procedure as in part a, with the only exception being the change from ≥ 0 to > 0.

3.2.1 $\Rightarrow$

For now, we assume that A is PD, which means for all $x \in \mathbb{R}^d$, $x^T Ax > 0$. Additionally, we know that

$$Au_i = \lambda_i u_i \tag{39}$$

For each eigenvalue $diag(\lambda_1, \dots, \lambda_d)$. Multiplying each side by u_i^T we get

$$\begin{aligned} u_i^T Au_i &= u_i^T \lambda_i u_i && \text{since } \lambda_i \text{ is a scalar} \quad (40)\\ u_i^T Au_i &= u_i^T u_i \lambda_i = \lambda_i && (41) \end{aligned}$$

So we get the important relation $u_i^T Au_i = \lambda_i$, and we know, from the assumption A is PD, that $u_i^T Au_i > 0$ for any u_i. Therefore, $0 < u_i^T Au_i = \lambda_i$ and $\lambda_i > 0$ in order for the relation $Au_i = \lambda_i u_i$ to hold. This will hold for any $i \in \{1, \dots, d\}$, and therefore we can say $\lambda_i > 0$ for $i \in \{1, \dots, d\}$ and the theorem is proved.

3.2.2 $\Leftarrow$

Now, we assume that $\lambda_i > 0$ for each i, and must prove that $x^T Ax > 0$. To complete this proof, take any vector x and define it as $x = U\tilde{x}$. Since $UU^T = I$, we know U has an inverse and therefore is full rank. Therefore, $x = U\tilde{x}$ will not suffer any loss of rank and can span $\mathbb{R}^{d\times 1}$. Using this, it is easy to see:

$$\begin{aligned} x^T Ax &= \tilde{x}^T U^T AU\tilde{x} && \text{from eigendecomposition know } U^T AU = \Lambda \quad (42)\\ &= \tilde{x}^T \Lambda \tilde{x} && (43) \end{aligned}$$

Since we know $\Lambda > 0$, by definition of a positive semi-definite matrix $\tilde{x}^T \Lambda \tilde{x} > 0$ and therefore

$$x^T Ax = \tilde{x}^T \Lambda \tilde{x} > 0 \tag{44}$$

and therefore, via the equality $x^T Ax > 0$ and the proof is complete.

4 Problem 4

For this problem, we are given iid Poisson random variables $\{X_1, \dots, X_n\}$ with intensity parameter λ and asked to determine the maximum likelihood estimator of λ. The maximum likelihood estimate is defined as

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log\, f(X_i; \lambda) \tag{45}$$

Where the Poisson distribution $f(X_i; \lambda)$ is

$$f(X_i; \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \tag{46}$$

Therefore, the log sum would be

$$\begin{aligned} \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) &= \sum_{i=1}^{n} \left(x_i \log(\lambda) - \lambda - \log(x_i!)\right) && (47)\\ &= \sum_{i=1}^{n} x_i \log(\lambda) - n\lambda - \sum_{i=1}^{n} \log(x_i!) && (48) \end{aligned}$$

To find the maximum, we find $\hat{\lambda}$ such that $\delta\left(\sum_{i=1}^{n} \log\, f(X_i;\lambda)\right)/\delta\hat{\lambda} = 0$.

$$= \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} - n = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{49}$$

And therefore $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the maximum likelihood estimator.

5 Problem 5

5.1 Part a

We are asked to show that is if f is strictly convex, then f has at most one global minimizer. To show this, I will use the constraint:

$$f(x+y) \geq f(x) + \bigtriangledown_x f(x) y \tag{50}$$

If f is strictly convex, this means that this constraint holds for every $x, y \in \mathbb{R}$. I will prove this by contradiction. Assume we have two global minimums x_1, x_2 in our strictly convex function. Since our set of $x \in \mathbb{R}$ is the entire real domain, we know at the minimum, we know that $\bigtriangledown_x f(x) = 0$. I.e., it is not a closed set of points but rather the entire domain and the minimum won't occur at the boundary. Therefore, at the minimum points x_1 and x_2

$$f(x_1 + y_1) \geq f(x_1) + 0 * y_1 \tag{51}$$
$$f(x_2 + y_2) \geq f(x_2) + 0 * y_2 \tag{52}$$

For any $y_1, y_2 \in \mathbb{R}$. For arguments sake, set $y_1 = x_2 - x_1$ and $y_2 = x_1 - x_2$ and the above become:

$$f(x_2) \geq f(x_1) \tag{53}$$
$$f(x_1) \geq f(x_2) \tag{54}$$

And the only way the above can hold is if $x_1 = x_2$ and therefore there can only be one global minimizer.

5.2 Part b

For the next two parts, we can use the following facts. The first of which is that a twice continually differentiable function admits the quadratic expansion

$$f(x) = f(y) + \langle \bigtriangledown f(y), x - y\rangle + \frac{1}{2}\langle x - y, \bigtriangledown^2 f(y)(x-y)\rangle + \sigma(\|x-y\|^2) \tag{55}$$

where $\sigma(t)$ denotes a function satisfying $\lim_{t\to 0} \frac{\sigma(t)}{t} = 0$, as well as the expansion

$$f(x) = f(y) + \langle \bigtriangledown f(y), x - y\rangle + \frac{1}{2}\langle x - y, \bigtriangledown^2 f(y + t(x-y))(x-y)\rangle \tag{56}$$

for some $t \in (0, 1)$.

For this part of the problem, I will use Equation (55) to prove this fact. Let's say that our local minimum is at $x^* = y$, and therefore

$$f(x) = f(x^*) + \langle \triangledown f(x^*), x - x^* \rangle + \frac{1}{2}\langle x - x^*, \triangledown^2 f(x^*)(x - x^*) \rangle + \sigma(\|x - x^*\|^2) \tag{57}$$

At a local minimum, we know that $\triangledown f(x^*) = 0$ and therefore the above can be simplified to:

$$f(x) - f(x^*) = \frac{1}{2}\langle x - x^*, \triangledown^2 f(x^*)(x - x^*) \rangle + \sigma(\|x - x^*\|^2) \tag{58}$$

By definition, any point around the local minimum must be larger than the minimum, i.e.,

$$f(x) - f(x^*) > 0 \tag{59}$$

Let us now consider the case when $x \to x^*$, or our reference point x starts to approach the local minimum. From definition, we know $\lim\limits_{t \to 0} \frac{\sigma(t)}{t} = 0$, or the σ function divided by t approaches zero. In the above equation, we simply have $\sigma(\|x - x^*\|^2)$ without the division of $\|x - x^*\|^2$. This means that this function will approach 0 faster as $x \to x^*$. Therefore, as $x \to x^*$, σ becomes very small or even approaches zero and will not factor much into the equation and we can say:

$$0 < f(x) - f(x^*) \approx \frac{1}{2}\langle x - x^*, \triangledown^2 f(x^*)(x - x^*) \rangle \tag{60}$$

$$\Rightarrow 0 < \langle x - x^*, \triangledown^2 f(x^*)(x - x^*) \rangle \tag{61}$$

$$0 < (x - x^*)^T \triangledown^2 f(x^*)(x - x^*) \tag{62}$$

And therefore $\triangledown^2 f(x^*)$ must be positive definite (by the definition of a positive definite matrix).

5.3 Part c

For this part, I consider Equation (56), which states

$$f(x) = f(y) + \langle \triangledown f(y), x - y \rangle + \frac{1}{2}\langle x - y, \triangledown^2 f(y + t(x - y))(x - y) \rangle \tag{63}$$

Additionally, from convex functions we know that, for any $x, y \in \mathbb{R}^d$,

$$f(x + y) \geq f(y) + \triangledown f(y)^T x \tag{64}$$

$$\Rightarrow f(x + y) - f(y) - \triangledown f(y)^T x \geq 0 \tag{65}$$

The constraint above can be simplified to

$$f(x) = f(y) + \triangledown f(y)^T (x - y) + \frac{1}{2}(x - y)^T \triangledown^2 f\left(y + t(x - y)\right)(x - y) \tag{66}$$

For our purposes, say we define $x = \bar{x} + y$. The above equation then becomes:

$$f(\bar{x} + y) = f(y) + \triangledown f(y)^T (\bar{x}) + \frac{1}{2}\bar{x}^T \triangledown^2 f\left(y + t\bar{x}\right)\bar{x} \tag{67}$$

or equivalently

$$f(\bar{x} + y) - f(y) - \triangledown f(y)^T (\bar{x}) = \frac{1}{2}\bar{x}^T \triangledown^2 f\left(y + t\bar{x}\right)\bar{x} \tag{68}$$

And in order for this function to be convex, we need

$$0 \leq (\bar{x} + y) - f(y) - \triangledown f(y)^T(\bar{x}) = \frac{1}{2}\bar{x}^T \triangledown^2 f\,(y + t\bar{x})\,\bar{x} \tag{69}$$

$$\Rightarrow 0 \leq \frac{1}{2}\bar{x}^T \triangledown^2 f\,(y + t\bar{x})\,\bar{x} \tag{70}$$

And therefore, I have shown that the Hessian needs to be positive definite for any $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$. As a note, the expression $y + t\bar{x}$ will cover all of $\mathbb{R}^d$ because $\bar{x}, y$ spans then entire $\mathbb{R}^d$.

5.4 Part d

We are given the function $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, where A is symmetric. We are asked to derive the Hessian of f, which would be derived as

$$f(x) = \tfrac{1}{2}x^T Ax + b^T x + c \tag{71}$$

$$\Rightarrow \triangledown_x f(x) = \tfrac{1}{2}Ax + \tfrac{1}{2}Ax + b^T = Ax + b^T \tag{72}$$

$$\Rightarrow \triangledown_x^2 f(x) = A \tag{73}$$

From the above sections, we know that a positive semi-definite Hessian implies convexity. It can also be said that a positive definite Hessian implies strict convexity, since there are no points which $\triangledown_x^2 f(x) = 0$ and therefore the convexity conditions will always strictly hold. Therefore, we can say

- $A \succeq 0 \Rightarrow f$ is convex
- $A \succ 0 \Rightarrow f$ is strictly convex