EE 127/EE 227AT– Spring 2018— Homework 5 Solutions

Michael Tong– SID: 24260171

Intro

So tried to do a total of 5 problems, managed to finish 3 and half finish the other 2.
Finished:
9.3, 9.5, 9.6
Attempted:
9.2, 9.7
Did not attempt:
9.4

Feel free to make suggestions to the solutions or leave other feedback.

9.2

(a) Since from 9.10, $\log(\det(\Theta)) = -\infty$ if Θ is not PD. Will assume that Θ is PD.

$\frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}_{\Theta}(x_i)$

From the book (9.8) $\mathbb{P}_{\gamma,\Theta}(x) = \exp[\sum_{s=1}^{p} \gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t - A(\Theta)]$
Since $A(\Theta) = -\frac{1}{2}\log\det(\Theta/2\pi)$, can rewrite this as
$\mathbb{P}_{\gamma,\Theta}(x) = \exp[\sum_{s=1}^{p} \gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t + \frac{1}{2}\log\det(\Theta/2\pi)]$
Since we are only concerned with $\mathbb{P}_{\Theta}(x)$, drop the γ term to get
$\mathbb{P}_{\Theta}(x) = \exp[-\frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t + \frac{1}{2}\log\det(\Theta/2\pi)]$

From this, taking the log you get
$\log \mathbb{P}_{\Theta}(x) = -\frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t + \frac{1}{2}\log\det(\Theta/2\pi)$
$\log \mathbb{P}_{\Theta}(x) = -\frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t + \frac{1}{2}\log\det(\Theta) - \frac{p}{2}(\log 2\pi)$
$\rightarrow$ Using the property that the trace of of a product can be written as the sum of entry-wise products of elements, i.e. $tr(X^\top Y) = \sum_{ij} X_{ij} Y_{ij}$.
$\rightarrow$ Let $\Theta = X^\top, x^\top x = Y$, then $tr(\Theta x^\top x) = \sum_{s,t=1}^{p} \theta_{st} x_s x_t$
$\log \mathbb{P}_{\Theta}(x) = -\frac{1}{2}tr(\Theta x^\top x) + \frac{1}{2}\log\det(\Theta) - \frac{p}{2}(\log 2\pi)$

So now can evaluate the sum
$\frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}_{\Theta}(x_i)$
$= \frac{1}{2}\log\det(\Theta) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N} tr(\Theta x^\top x)$
$= \frac{1}{2}\log\det(\Theta) - \frac{p}{2}\log(2\pi) - \frac{1}{2}tr(\Theta \frac{1}{N}\sum_{i=1}^{N} x_i^\top x_i)$
$\rightarrow$ substitute $S = \frac{1}{N}\sum_{i=1}^{N} x_i^\top x_i)$
$= \frac{1}{2}\log\det(\Theta) - \frac{p}{2}\log(2\pi) - \frac{1}{2}tr(\Theta S)$
$\rightarrow$ scale the equation by 2
$= \log\det(\Theta) - p\log(2\pi) - tr(\Theta S)$
$\rightarrow$ substitute $p\log(2\pi) = C$
$= \log\det(\Theta) - tr(\Theta S) - p\log(2\pi)$
Which the expression you were supposed to get.
Also note that $C = p\log(2\pi)$ does not depend on Θ

(b) The sign of the suggested solution of $\nabla f(\Theta) = \Theta^{-1}$ is wrong. To prove this, consider the 1x1 matrix Θ.
Then $f(\Theta) = -\log\det\Theta = -\log(\Theta)$ and $\nabla f(\Theta) = -\Theta^{-1}$
So will instead prove that $\nabla f(\Theta) = -\Theta^{-1}$

$f(\Theta) = -\log\det\Theta$
$\nabla f(\Theta) = -\frac{1}{\det\Theta}\nabla(\det\Theta)$
$\nabla f(\Theta) = -\frac{1}{\det\Theta}\nabla(\det\Theta)$
$\nabla f(\Theta) = -\frac{1}{\det\Theta}adj(\Theta)$, where $adj(\Theta)$ is is the adjugate of Θ
Since for an invertible matrix, $\frac{1}{\det\Theta}adj(\Theta) = \Theta^{-1}$
$\nabla f(\Theta) = -\Theta^{-1}$
Thus have proved that $\nabla f(\Theta) = -\Theta^{-1}$

$\nabla^2 f(\Theta) = -\nabla\Theta^{-1}$

Since Θ is PD, so is Θ^{-1}
Let Θ^{-1} have an eigenvalue decomposition of $\Theta^{-1} = U\Sigma^{-1}U^\top$
$\nabla^2 f(\Theta) = -\nabla U\Sigma^{-1}U^\top$

The derivative of the inverse of a matrix can be expressed as
$\frac{\partial Y^{-1}}{\partial x} = -Y^{-1}\frac{\partial Y}{\partial x}Y^{-1}$
From this, you can write that for any eigenvalue of Θu_i
$\frac{\partial \Theta^{-1}}{\partial u_i} = -\Theta^{-1}(\frac{\partial \Theta}{\partial u_i})\Theta^{-1}$
$\frac{\partial \Theta^{-1}}{\partial u_i} = -\Theta^{-2}(u_i\lambda_i^{-1})$
$\frac{\partial \Theta^{-1}}{\partial u_i} = \Theta^{-2}(u_i\lambda_i^{-2})$
Since Θ and Θ^{-1} are PD, then this expression is positive for each eigenvector u_i
Since the function is convex if you move in the direction of each of the eigenvectors, it is convex for the entire space and $f(\Theta)$ is convex

(c) Since the function is convex, the maximum is obtained when $\nabla f(\Theta) = 0$
$\nabla(\log\det\Theta - tr(S\Theta)) = \Theta^{-1} - \nabla tr(S\Theta)$
$0 = \Theta^{-1} - S$
$S = \Theta^{-1}$
When $N > p$, then Θ is not invertable and the MLE solution does not exist.

(d) With l1-regularization, the optimization problem becomes the expression in (9.13)
$\hat{\Theta} \in \arg\max \log\det\Theta + tr(S\Theta) - \lambda\rho_1(\Theta)$
Since the problem is convex, at the optimal solution you have the equality given in (9.14)
$\Theta^{-1} - S - \lambda\cdot\Psi = 0$
$W - S - \lambda\cdot\Psi = 0$
Where $\psi_{jj} = 0, \psi_{jk} = sign(\theta_{jk})$ if $\theta_{jk} \neq 0$ else $\psi_{jk} \in [-1, 1]$ if $\theta_{jk=0}$

Thus we can write this as an optimization problem
$\hat{\Theta} \in \arg\max_{\Theta,W} \log\det\Theta + tr(S\Theta) - \lambda\rho_1(\Theta)$
such that
$W\Theta = I$
$w_{ii} - s_{ii} = 0$
$|w_{ij} - s_{ij}| \leq \lambda, w_{ij} = 0, i \neq j$
$w_{ij} - s_{ij} = \lambda sign(\theta_{ij}), w_{ij} \neq 0, i \neq j$

Thus the you can write KarushKuhnTucker equations as follows:
There exist some

9.3

Looking at the subgradient equation in (9.13), the graphical lasso algorithm has a gradient of
$\Theta^{-1} - S - \lambda \cdot \Psi = 0$
$\rightarrow$ if $\lambda = 0$, then the equation becomes
$\Theta^{-1} - S = 0$
$S = \Theta^{-1}$
Since Θ symmetric, you can write it's eigenvalue decomposition as $U\Sigma U^\top$, where Σ is a diagonal matrix.
Since Θ is PD, then you know that all of it's eigenvalues are > 0, and the inverse exists and is unique. Thus you can take the inverse and it gives a unique solution of $S = \Theta^{-1}$

9.5

a) From (9.27)
$\hat{\theta}^s \in \arg\min_{\theta^s \mathbb{R}^p} = \frac{1}{N}\sum_{i=1}^N l[x_{is}, \eta_{\theta^s}(x_{i,\setminus\{s\}})] + \lambda_{t\in V\setminus\{s\}}|\theta_{st}|$
Since θ is given in the expression $\mathbb{P}(x_s|x_{V\setminus\{s\}};\theta)$, you can ignore the regularization term and the optimization problem becomes
$\hat{\theta}^s \in \arg\min_{\theta^s \mathbb{R}^p} = \frac{1}{N}\sum_{i=1}^N l[x_{is}, \eta_{\theta^s}(x_{i,\setminus\{s\}})]$
Which is the standard logistic regression problem.

b) The function l is the negative log-likelihood function for the binomial distribution
Now we will solve the follwing minimization problem
$\min_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = 0$

$\nabla_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = \nabla_y - \frac{1}{N}\sum_{i=1}^N (1-x_i)\log(1-\sigma(y)) + x_i \log(\sigma(y))$
where σ is the sigmoid function $\sigma(y) = \frac{1}{1+e^{-y}}$
$\nabla_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = -\frac{1}{N}\sum_{i=1}^N (1-x_i)\nabla_y[\log(1-\sigma(y))] + x_i\nabla_y[\log(\sigma(y))]$

$\nabla_y[\log(\sigma(y))] = \frac{1}{\sigma(y)}\nabla_y \sigma(y)$
$\nabla_y[\log(\sigma(y))] = \frac{1}{\sigma(y)}\nabla_y (1+e^{-y})^{-1}$
$\nabla_y[\log(\sigma(y))] = \frac{1}{\sigma(y)}(-1)(1+e^{-y})^{-2}(-e^{-y})$
$\nabla_y[\log(\sigma(y))] = \frac{1}{\sigma(y)}(1+e^{-y})^{-2}(e^{-y})$
$\nabla_y[\log(\sigma(y))] = \frac{1}{\sigma(y)}(\sigma(y)(1-\sigma(y))$
$\nabla_y[\log(\sigma(y))] = 1-\sigma(y)$

$\nabla_y[\log(1-\sigma(y))] = \nabla_y[\log(\sigma(-y))]$
$\nabla_y[\log(1-\sigma(y))] = -(1-\sigma(-y))$
$\nabla_y[\log(1-\sigma(y))] = \sigma(-y)-1$
$\nabla_y[\log(1-\sigma(y))] = (1-\sigma(y))-1$
$\nabla_y[\log(1-\sigma(y))] = -\sigma(y)$

$\nabla_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = -\frac{1}{N}\sum_{i=1}^N (1-x_i)\nabla_y[1-\log(\sigma(y))] + x_i\nabla_y[\log(\sigma(y))]$
$\nabla_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = -\frac{1}{N}\sum_{i=1}^N (1-x_i)(-\sigma(y)) + x_i(1-\sigma(y))$
$\nabla_y \frac{1}{N}\sum_{i=1}^N l[x_i, y] = -\frac{1}{N}\sum_{i=1}^N -\sigma(y) + x_i\sigma(y) + x_i - x_i\sigma(y)$
$\nabla_y \frac{1}{N} - \sum_{i=1}^N l[x_i, y] = \frac{1}{N}\sum_{i=1}^N \sigma(y) - x_i$
Now set the derivative equal to 0 and solve
$\sigma(y) = \frac{1}{N}\sum_{i=1}^N x_i$
$\sigma(y) = \bar{x_i}$

From this, in order for the estimator to be Fischer-consistent, as long as $\frac{1}{N}\sum_{i=1}^N x_{is}$ approaches the true mean $\bar{x_{is}}$ as N approaches infinity there exists some $\eta_{\theta^s}(x_{i\setminus\{s\}})$ such that $\sigma(y) = \eta_{\theta^s}(x_{i\setminus\{s\}}) = \bar{x_s}$ which is the true conditional distribution. Since estimating the mean by taking the average of samples is Fischer-consistent, the logistic regression problem is also Fischer-consistent and the true conditional distribution is the population minimizer.

9.6

1. From (9.38)
   $\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^\top & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0^\top & 1 \end{pmatrix}$
   From this, equation, we know that
   $W_{11}\theta_{12} + w_{12}\theta_{22} = 0$
   $w_{12}\theta_{22} = -W_{11}\theta_{12}$
   $w_{12} = -W_{11}\theta_{12}/\theta_{22}$
   $\rightarrow$ let $\beta = -\theta_{12}/\theta_{22}$ $w_{12} = W_{11}\beta$

   Since Θ represents the inverse of a covariance matrix, you know that Θ is PD. Since Θ is PD, all elements along the diagonal of Θ are positive numbers. Thus $\theta_{22} > 0$

   From (9.37)
   $w_{12} - s_{12} - \lambda \cdot sign(\theta_{12}) = 0$
   $\rightarrow$ substituting $w_{12} = W_{11}\beta$
   $W_{11}\beta - s_{12} - \lambda \cdot sign(\theta_{12}) = 0$
   $\rightarrow$ substituting $\beta = -\theta_{12}/\theta_{22}$
   $W_{11}\beta - s_{12} - \lambda \cdot sign(-\beta\theta_{22}) = 0$
   $W_{11}\beta - s_{12} + \lambda \cdot sign(\beta\theta_{22}) = 0$
   $\rightarrow$ since $\theta_{22} > 0$
   $W_{11}\beta - s_{12} + \lambda \cdot sign(\beta) = 0$
   Which is the expression you were supposed to get

9.7

a) From (9.38)

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^\top & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0^\top & 1 \end{pmatrix}$$

The block inverse of a matrix can be written as

$$\begin{pmatrix} A & B \\ B^\top & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - B^\top A^{-1}B)^{-1}B^\top A^{-1} & -A^{-1}B(D - B^\top A^{-1}B)^{-1} \\ -(D - B^\top A^{-1}B)^{-1}B^\top A^{-1} & (D - B^\top A^{-1}B)^{-1} \end{pmatrix}$$

And the matrix exists as long as $(D - B^\top A^{-1}B)^{-1}$ is non-singular.

Since $\begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}^{-1} = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^\top & w_{22} \end{pmatrix}$. If we let $A = \Theta_{11}, B = \theta_{12}, D = \theta_{22}$, then

$w_{12} = -A^{-1}B(D - B^\top A^{-1}B)^{-1}$

$\rightarrow$ substitute $w_{22} = (D - B^\top A^{-1}B)^{-1}$

$w_{12} = -A^{-1}Bw_{22}$

$w_{12} = -\Theta_{11}^{-1}\theta_{12}w_{22}$

So the expression (9.37)

$w_{12} - s_{12} - \lambda \cdot sign(\theta_{12}) = 0$

Can be rewritten as

$-\Theta_{11}^{-1}\theta_{12}w_{22} - s_{12} - \lambda \cdot sign(\theta_{12}) = 0$

$\Theta_{11}^{-1}\theta_{12}w_{22} + s_{12} + \lambda \cdot sign(\theta_{12}) = 0$

Which is the expression you are supposed to get.

b) In order to update $\hat{\Theta}$, just update θ_{12}. Since there are $O(p)$ elements in θ_{12}, this operation takes $O(p)$ time.

Now to calculate the updated W

Let θ_{12} be the original θ_{12}, θ'_{12} be the updated value, and $\delta_{12} = \theta'_{12} - \theta_{12}$.

Using the Sherman-Morrison formula of

$(A + uv^\top)^{-1} = A^{-1} + \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}$

Since $\hat{\Theta}^{-1} = W$ can write that.

$$(\Theta + \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top)^{-1} = W - \frac{W \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top W}{1 + \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top W \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix}}$$

Then let $\Theta' = \Theta + \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top, W' = W - \dfrac{W \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top W}{1 + \begin{pmatrix} 0 \\ 1 \end{pmatrix}^\top W \begin{pmatrix} \delta_{12} \\ 0 \end{pmatrix}}$

To get $\Theta'^{-1} = W'$

Then we can also write

$$(\Theta' + \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top)^{-1} = W' - \frac{W'\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top W'}{1+\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top W'\begin{pmatrix}0\\1\end{pmatrix}}$$

Then let $\Theta'' = \Theta' + \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top, W'' = W' - \frac{W'\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top W'}{1+\begin{pmatrix}\delta_{12}\\0\end{pmatrix}^\top W'\begin{pmatrix}0\\1\end{pmatrix}}$

To get $\Theta''^{-1} = W''$

So to calculate the updated W' from W, first calculate $W\begin{pmatrix}\delta_{12}\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}^\top W$. Since our are doing matrix-vector multiplication, this takes $O(p^2)$ time.
Then with the two vectors in the numerator take the outer product, this also takes $O(p^2)$ time.
Looking at the expression of the denominator, it takes $O(p^2)$ time to calculate as well.
Finally, to calculate W' given W and the fraction expression, subtracting a matrix takes $O(p^2)$ time.
Thus to calculate W' from W takes $O(p^2)$ time.
Similarly, you can calculate W'' from W' in $O(p^2)$ time, and your updated W is W''.

c) To move to a new block of equations in $O(p^2)$ we will write the expression in the form of (9.42)
From excercise 9.6, we proved that $w_{12} = -W_{11}\theta_{12}/\theta_{22}$
From part a), we showed that $w_{12} = -\Theta_{11}^{-1}\theta_{12}w_{22}$
Thus $-\Theta_{11}^{-1}w_{22} = -W_{11}/\theta_{22}$
So we can rewrite the expression (9.42)
$\Theta_{11}^{-1}\theta_{12}w_{22} + s_{12} + \lambda \cdot sign(\theta_{12})$ as
$(W_{11}/\theta_{22})\theta_{12} + s_{12} + \lambda \cdot sign(\theta_{12})$
The only thing we need to calculate to set up this new batch of equations is W_{11}/θ_{22} which takes $O(p^2)$ time.
Thus we can move to a new block of equations in $O(p^2)$ time.

d) Algorithm is below

Result: Estimate of $\hat{\Theta}$
calculate S to be the sampled covariance of your data;
initialize $\hat{\Theta}, W$ to be diagonal matrices where $\hat{\Theta} = diag(S), W = \hat{\Theta}^{-1}$;
while *The estimate of* $\hat{\Theta}$ *has not converged* **do**
    Pick a random instructions;
    if *condition* **then**
        instructions1;
        instructions2;
    else
        instructions3;
    end
end

Algorithm 1: Primal graphical lasso algorithm