708 HW2

Shawn Lyu

1

1.1

(a)

We can easily use Jensen's inequality to prove the inequality. Let $Y = \frac{q(x)}{p(x)}$ be a random variable, we know that the function $\phi(y) = -\log(y)$ is a convex function. We know by properties of logarithm that $\phi(y) = \log\left(\frac{p(x)}{q(x)}\right)$

$$\begin{aligned}\int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx &= \mathbb{E}[\phi(Y)] \geq \phi(\mathbb{E}[Y]) \\ &= -\log\left(\int p(x)\frac{q(x)}{p(x)}dx\right) \\ &= -\log\left(\int p(x)dx\right) = 0\end{aligned}$$

Therefore $D_{KL}(p(x)||q(x)) \geq 0$. When $p(x) = q(x)$, the KL divergence just becomes

$$\int p(x)\log 1 dx = \int 0 dx = 0$$

(b)

By the Strong Law of Large Numbers we know that as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N}\log(p(x^{(i)};\theta)) \to \mathbb{E}[\log(p(x;\theta))]$$

almost surely, so as $N \to \infty$,

$$\hat{\theta}_{MLE} \to \arg\max_{\theta}\mathbb{E}[\log(p(x;\theta))]$$

almost surely as well.

Note that by definition,

$$\begin{aligned}\arg\min_\theta \int p(x;\theta^*)\log\left(\frac{p(x;\theta^*)}{p(x;\theta)}\right)dx &= \arg\min_\theta -\int p(x;\theta^*)\log\left(\frac{p(x;\theta)}{p(x;\theta^*)}\right)dx \\ &= \arg\max_\theta \mathbb{E}[\log\left(\frac{p(x;\theta)}{p(x;\theta^*)}\right)] \\ &= \arg\max_\theta \mathbb{E}[\log p(x;\theta) - \log p(x;\theta^*)] \\ &= \arg\max_\theta \mathbb{E}[\log p(x;\theta)] - \mathbb{E}[\log p(x;\theta^*)] \\ &= \arg\max_\theta \mathbb{E}[\log p(x;\theta)]\end{aligned}$$

Therefore, we know that as $N \to \infty$,

$$\hat{\theta}_{MLE} \to \arg\min_\theta D_{KL}(p(x;\theta)||p(x;\theta^*))$$

almost surely.

(c)

Proof: By simply marginalizing $p(x_A, x_B), q(x_A, x_B)$ over x_B, we know that $p(x_A) = \sum_{x_B} p(x_A, x_B), q(x_A) = \sum_{x_B} q(x_A, x_B)$. Focus on the second term of the right hand side first and assuming a fixed $x_A = x_a$, we have

$$\begin{aligned}D_{KL}(p(x_B|x_a)||q(x_B|x_a)) &= \int_{x_B} \frac{p(x_a,x_B)}{p(x_a)}\log\left(\frac{p(x_a,x_B)/p(x_a)}{q(x_a,x_B)/q(x_a)}\right)dx \\ &= \int_{x_B} \frac{p(x_a,x_B)}{p(x_a)}\left(\log(\frac{p(x_a,x_B)}{q(x_a,x_B)}) - \log(\frac{p(x_a)}{q(x_a)})\right)dx\end{aligned}$$

Therefore we know that for some fixed $x_A = x_a$ and with $p(x_a)$ given

$$\begin{aligned}&p(x_a)D_{KL}(p(x_B|x_a)||q(x_B|x_a)) \\ &= \int_{x_B} p(x_a,x_B)\left(\log(\frac{p(x_a,x_B)}{q(x_a,x_B)}) - \log(\frac{p(x_a)}{q(x_a)})\right)dx \\ &= \int_{x_B} p(x_a,x_B)\log(\frac{p(x_a,x_B)}{q(x_a,x_B)})dx - \int_{x_B} p(x_a,x_B)\log(\frac{p(x_a)}{q(x_a)})dx \\ &= \int_{x_B} p(x_a,x_B)\log(\frac{p(x_a,x_B)}{q(x_a,x_B)})dx - p(x_a)\log(\frac{p(x_a)}{q(x_a)}dx\end{aligned}$$

, as x_B gets marginalized over the integral.

When summing over the support of $p(x_A)$, we know that the first term is just

$$\int_{x_A,x_B} p(x_A,x_B)\log(\frac{p(x_A,x_B)}{q(x_A,x_B)})dx = D_{KL}(p(x_A,x_B)||q(x_A,x_B))$$

and the second term becomes

$$\int_{x_A} p(x_A) \log(\frac{p(x_A)}{q(x_A)} dx = D_{KL}(p(x_A)||q(x_A))$$

Therefore, we know that

$$\begin{aligned}
&D_{KL}(p(x_A, x_B)||q(x_A, x_B)) \\
=&D_{KL}(p(x_A)||q(x_A)) + D_{KL}(p(x_A, x_B)||q(x_A, x_B)) - D_{KL}(p(x_A)||q(x_A)) \\
=&D_{KL}(p(x_A)||q(x_A)) + \sum_{x_A} p(x_A) D_{KL}(p(x_B|x_a)||q(x_B|x_a))
\end{aligned}$$

(d)

Proof: Fix on a certain cluster C_k and refer to it as C for both convenience and consistency with the writeup. Let ℓ be the log likelihood of the model. To show that IPF is equivalent to coordinate ascent, we only need to prove that

$$\phi_C^{(t)}(x_C) \frac{\epsilon(x_C)}{p^{(t)}(x_C)} = \arg\max_{x_C} \ell(x)$$

From (b) we also know that as $N \to \infty$, $\arg\max_{x_C} \ell(x)$ converges almost surely to the set of values that minimize the KL divergence between the two distributions. Let p^* be the true distribution of the model, we may write the KL divergence as

$$D_{KL}(p^*(x_{U \setminus C}, x_C)||p(x_{U \setminus C}, x_C|\theta))$$

, which, according to (c), is just

$$D_{KL}(p^*(x_C)||p(x_C|\theta)) + \sum_{x_C} p^*(x_C) D_{KL}(p^*(x_{U \setminus C}|x_C)||p(x_{U \setminus C}|x_C, \theta))$$

We shall first prove that the second term is completely unaffected by changes in $\phi_C(x_C)$. We know that

$$\begin{aligned}
p(x_{U \setminus C}|x_C, \theta) &= \frac{p(x_{U \setminus c}, x_C|\theta)}{p(x_C|\theta)} \\
&= \frac{\prod_{i \neq k} \phi_{C_i}^{(t)}(x_{C_i}) \phi_C^{(t+1)}(x_C)/Z}{\sum_{x_{U \setminus C}} \prod_{i \neq k} \phi_{C_i}^{(t)}(x_{C_i}) \phi_C^{(t+1)}(x_C)/Z} \\
&= \frac{\prod_{i \neq k} \phi_{C_i}^{(t)}(x_{C_i})}{\sum_{x_{U \setminus C}} \prod_{i \neq k} \phi_{C_i}^{(t)}(x_{C_i})}
\end{aligned}$$

, where $Z = \sum_x \prod_{i \neq k} \phi_{C_i}^{(t)}(x_{C_i}) \phi_C^{(t+1)}(x_C)$. Since p^*, the true distribution, can never be changed, we know that changing $\phi_C(x_C)$ will not change second term at all.

We then show that $\phi_C^{(t+1)}(x_C)$ minimizes the first term, $D_{KL}(p^*(x_C)||p(x_C|\theta))$. Note that, by (b) again, this is equivalent to maximizing the likelihood with other parameters in the model fixed. By definition of the IPF steps, this is exactly why we set

$$\phi_C^{(t+1)}(x_C) \leftarrow \phi_C^{(t)}(x_C)\frac{\epsilon(x_C)}{p^{(t)}(x_C)}$$

.

Therefore, the update minimizes the KL divergence between the two distributions, and the algorithm can be viewed as coordinate ascent.

1.2

(a)

For the sake of convenience, we take out one arbitrary element from the two matrix representations, and we know that

$$[\mathcal{I}(\theta)]_{ij} = \mathbb{E}[\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)]$$

and we want to show that this is equivalent to

$$-\mathbb{E}[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x;\theta)]$$

. First of all, observe that

$$\frac{\partial}{\partial\theta_i}\log p(x;\theta) = \frac{\frac{\partial}{\partial\theta_i}p(x;\theta)}{p(x;\theta)}$$

. As a result we know that

$$\begin{aligned}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x;\theta) &= \frac{\partial}{\partial\theta_j}\frac{\frac{\partial}{\partial\theta_i}p(x;\theta)}{p(x;\theta)} = \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)}{p(x;\theta)} - \frac{\frac{\partial}{\partial\theta_i}p(x;\theta)\frac{\partial}{\partial\theta_j}p(x;\theta)}{p^2(x;\theta)} \\ &= \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)}{p(x;\theta)} - \frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)\end{aligned}$$

We first find the expected value of the first term, and realize that under regularity assumptions

$$\mathbb{E}[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)}{p(x;\theta)}] = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathbb{E}[\frac{p(x;\theta)}{p(x;\theta)}] = \frac{\partial^2}{\partial\theta_i\partial\theta_j}1 = 0$$

as a result, we have

$$\mathbb{E}[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x;\theta)] = -\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta) = -\mathcal{I}(\theta)]_{ij}$$

, and thus the two representations are equivalent.

(b)

First observe that

$$\log p(x;\theta) = \log(h(x)) + \theta^T T(x) - A(\theta)$$

We then easily have

$$\nabla_\theta \mathbb{E}[\log p(x;\theta)] = \mathbb{E}[T(x)] - \nabla_\theta A(\theta)$$

From above we obtain the Fisher information matrix is just

$$-\nabla^2_\theta \mathbb{E}[\log p(x;\theta)] = \nabla^2_\theta A(\theta)$$

(c)

We first write out the KL divergence

$$D_{KL}(p(x;\theta)||p(x;\theta+\delta)) = \int p(x;\theta)(\log(p(x;\theta)) - \log(p(x;\theta+\delta)))dx$$

Assume δ is sufficiently small and view the above KL divergence as a function of $\delta + \theta$. By the regularity assumptions on $p(x;\theta)$ and for better representation let $\theta' = \delta + \theta$, the Taylor expansion for the above around θ becomes

$$\begin{aligned} D_{KL}(p(x;\theta)||p(x;\theta)) + \delta \int \frac{\partial}{\partial \theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta'))dx \\ + \frac{\delta^2}{2} \int \frac{\partial^2}{\partial^2 \theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta'))dx + o(\delta^2) \end{aligned}$$

From (a), it is obvious that the first term is 0. The function being integrated in the second term can be written as

$$-p(x;\theta)\left[\frac{\partial}{\partial \theta'} \log p(x;\theta')\right]_{\theta'=\theta} = -p(x;\theta)\frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)} = \frac{\partial}{\partial\theta}p(x;\theta)$$

Change the order of differentiation and integral we observe the second term becomes

$$\frac{\partial}{\partial\theta}\int p(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0$$

We are left with the third term. We easily note that

$$\frac{\partial^2}{\partial^2\theta'}p(x;\theta)(\log p(x;\theta) - \log p(x;\theta')) = -p(x;\theta)\frac{\partial^2}{\partial^2\theta'}\log p(x;\theta')$$

which just means that the integral in the third term can also be viewed as $-\mathbb{E}[\frac{\partial^2}{\partial^2\theta'}\log p(x;\theta')]$. We can then obtain the following that

$$\begin{aligned} D_{KL}(p(x;\theta)||p(x;\theta+\delta)) &= 0 + 0 - \frac{\delta^2}{2}\mathbb{E}[\frac{\partial^2}{\partial^2\theta'}\log p(x;\theta')] + o(\delta^2) \\ &= -\frac{\delta^2}{2}\mathbb{E}[\frac{\partial^2}{\partial^2\theta'}\log p(x;\theta')] + o(\delta^2) \end{aligned}$$

(d)

Similar to above, we can write the Taylor expansion around θ as

$$D_{KL}(p(x;\theta)||p(x;\theta))+\delta^T \int \nabla_{\theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta'))dx$$
$$+\frac{\delta^T}{2}(\int \nabla^2_{\theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta'))dx)\delta + o(\|\delta\|^2)$$

From (a) we know the first term is 0. We focus on the second term. Note that

$$\log p(x;\theta') = h(x) + \theta'^T T(x) - A(\theta')$$

, we have

$$\begin{aligned}\nabla_{\theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta')) &= -\nabla_{\theta'} p(x;\theta)(h(x) + \theta'^T T(x) - A(\theta')) \\ &= -p(x;\theta)T(x) + p(x;\theta)\left[\nabla_{\theta'} A(\theta')\right]_{\theta'=\theta}\end{aligned}$$

Consequently, when δ is small, the second term becomes

$$-\delta^T \mathbb{E}[T(x)] + \delta^T \mathbb{E}[\nabla_\theta A(\theta)] = -\delta^T \mathbb{E}[T(x) - \nabla_\theta A(\theta)]$$

At the same time, observe that

$$\nabla_\theta p(x;\theta) = p(x;\theta)(T(x) - \nabla_\theta A(\theta))$$

Which tells us

$$\begin{aligned}\mathbb{E}[T(x) - \nabla_\theta A(\theta)] &= \int p(x;\theta)\frac{\nabla_\theta p(x;\theta)}{p(x;\theta)}dx \\ &= \nabla_\theta \int p(x;\theta)dx = \nabla_\theta 1 = 0\end{aligned}$$

, and the second term is also 0. Focusing on the third term, we have

$$\begin{aligned}\nabla^2_{\theta'} p(x;\theta)(\log p(x;\theta) - \log p(x;\theta')) &= -\nabla^2_{\theta'} p(x;\theta)(h(x) + \theta'^T T(x) - A(\theta')) \\ &= -p(x;\theta)\nabla^2_\theta A(\theta)\end{aligned}$$

, and the integral evaluates to $\mathbb{E}[\nabla^2_\theta A(\theta)] = \nabla^2_\theta A(\theta)$.

Plug the results back into the Taylor expansion, we know that the KL divergence between the two distributions is just

$$0 + 0 + \frac{1}{2}\delta^T \nabla^2_\theta A(\theta)\delta + o(\|\delta\|^2) = \frac{1}{2}\delta^T \nabla^2_\theta A(\theta)\delta + o(\|\delta\|^2)$$

2

2.2

It can be easily observed from the equation that $\frac{\partial}{\partial \theta_i} \frac{\lambda}{2} \sum_{i=1}^{k} \theta_i^2 = \lambda \theta_i$ and that $\frac{\partial}{\partial \theta_i} \sum_{i=1}^{k} \theta_i f_i(Y, X) = f_i(Y, X) = E_D[f_i]$. We are then left with finding the partial derivative of $\log(Z_X(\theta))$.

With chain rule we have

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \log Z_X(\theta) &= \frac{1}{Z_X(\theta)} \frac{\partial}{\partial \theta_j} Z_X(\theta) \\
&= \frac{1}{Z_X(\theta)} \sum_{Y'} \exp(\sum_{i=1}^{k} \theta_i f_i(D_i)) f_j(D_j) \\
&= \sum_{Y'} (\frac{1}{Z_X(\theta)} \exp(\sum_{i=1}^{k} \theta_i f_i(D_i))) f_j(D_j) \\
&= \sum_{Y'} (\frac{1}{Z_X(\theta)} \exp(\sum_{i=1}^{k} \theta_i f_i(D_i))) f_j(Y', X) = E_\theta[f_j]
\end{aligned}$$

as the value of $f_j(D_j)$ can easily be obtained when Y' and X are given.

We then easily know that

$$\frac{\partial}{\partial \theta_i} \text{nll}(X, Y, \theta) = E_\theta[f_j] - E_D[f_i] + \lambda \theta_i$$

2.3

To explain the differences in (7), we only need to show that under this setting the second term can be expressed like this. In the first model, we summed over the products of all the θ_js with their respective $f_j \in \{f^{(i)}\}$. By the distributive property of multiplication, we easily know that this is equivalent to first summing over all the $f_j \in \{f^{(i)}\}$ and then multiply them with the shared parameter θ_j.

Since θ_j is now being shared by multiple f_js, by the addition rule

of differentiation, we should take the sum of $E_\theta[f_j]$ and $E_D[f_j]$ for all $f_j \in \{f^{(i)}\}$ when calculating the partial derivative of θ_i. The regularization term need not to be summed because it is only calculated once in our penalized negative log likelihood.

2.4

(a)

There will be 26 θ_c^C's: one for each possible character in the alphabet.

There will be 1664 $\theta_{c,d}^I$'s: there are 32 pixels in the picture for each character, and each pixel is binary. Thus there will be 64 parameters needed to represent

the picture for one character. There are 26 possible characters, so there are $26 \times 64 = 1664$ parameters needed for $\theta^I_{c,d}$.

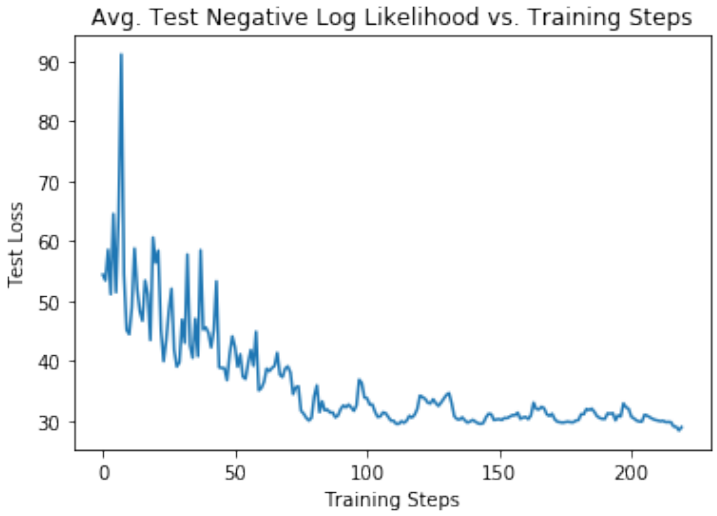
There will be 676 $\theta^P_{c,d}$'s: there are 26 possibilities for the character of Y_i, and 26 possibilities for Y_{i+1}. In total, there are $26 \times 26 = 676$ parameters needed.

In total, there are $26 + 1664 + 676 = 2366$ parameters, exactly the amount we need.

(b)

The negative log likelihood, sampleNLL, is 9.77428961406. The L_2 norm of the gradient is 9.62537942079.

2.5



3

3.1

1. Note that by definition of the model, $y \perp\!\!\!\perp x|\mu$, so by definition of conditional independence $p(y|\mu_{ij}, \sigma) = p(y|\mu_{ij}, x, \sigma)$. We easily know that

$$\begin{aligned} p(y|x,\theta) &= \int_{\mu} p(y,\mu|x,\theta)d\mu \\ &= \sum_i g_i(x,v_i) \sum_j g_{j|i}(x,v_{ij}) \int_{\mu_{ij}} p(y|\mu_{ij},x,\sigma)p(\mu_{ij}|x,\sigma)d\mu_{ij} \end{aligned}$$

Since we are essentially taking the marginal distribution over a multivariate Gaussian distribution, $p(y, \mu_{ij}|x, \sigma)$, the resulting distribution should also be multivariate Gaussian. Viewing the model as a linear Gaussian system, we know that

$$\mathbb{E}[y|x,\sigma] = \mathbb{E}[\mathbb{E}[y|\mu_{ij},x,\sigma]] = \mathbb{E}[\mu_{ij}] = \mathbb{E}[U_{ij}x] = U_{ij}x$$

. Let ϵ_i and ϵ denote the two sources of noise in the model, where we let $\mu_{ij} = U_{ij}x + \epsilon_i$ and $y = \mu_{ij} + \epsilon$. We then know that

$$\begin{aligned} Cov[y|x,\sigma] &= \mathbb{E}[(y - U_{ij}x)(y - U_{ij}x)^T|x,\sigma] \\ &= \mathbb{E}[(U_{ij}x + \epsilon + \epsilon_i - \mu_{ij})(U_{ij}x + \epsilon + \epsilon_i - \mu_{ij})^T|x,\sigma] \\ &= \mathbb{E}[\epsilon\epsilon^T + \epsilon_i\epsilon_i^T + \epsilon\epsilon_i^T + \epsilon_i\epsilon^T|x,\sigma] = (\sigma^2 + \sigma_i^2)I \end{aligned}$$

, as from the model we know that $\epsilon \perp\!\!\!\perp \epsilon_i$.

Therefore, we know that $y|x, \sigma \sim \mathcal{N}(U_{ij}x, \sigma^2 I)$, and the distribution can be written as

$$p(y|x,\theta) = \sum_i g_i(x,v_i) \sum_j g_{j|i}(x,v_{ij}) \frac{\exp(-\frac{1}{2(\sigma^2+\sigma_i^2)}(y - U_{ij}x)^T(y - U_{ij}x))}{(2\pi)^{n/2}(\sigma^2 + \sigma_i^2)^{n/2}}$$

2. By Bayes' rule we easily have

$$h_{j|i} = \frac{g_{j|i}P_{ij}(y|x,\theta_E)}{\sum_j g_{j|i}P_{ij}(y|x,\theta_E)}$$

3. Even though the final output may look like a linear combination of the features, the values of g and μ are both affected by x. Therefore, the terms μ_i and μ are not simple linear combination of the features, but are rather non-linear stochastic functions of x. As a result, our final model is also non-linear.

3.2

(a)

1. We take the expected value over $z_{ij}^{(t)}$ first, note that $z_{ij}^{(t)}$ is an indicator random variable, and realize that

$$\begin{aligned}
\mathbb{E}[z_{ij}^{(t)}|\mathcal{X}] &= P(z_{ij}^{(t)} = 1|y^{(t)}, x^{(t)}, \theta) \\
&= \frac{P(y^{(t)}|z_{ij}^{(t)} = 1, x^{(t)}, \theta)P(z_{ij}^{(t)} = 1|x^{(t)}, \theta)}{P(y^{(t)}|x^{(t)}, \theta)} \\
&= \frac{P(y^{(t)}|x^{(t)}, \theta_E)g_i^{(t)}g_{j|i}^{(t)}}{\sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P(y^{(t)}|x^{(t)}, \theta_E)} \\
&= \frac{P_{ij}^{(t)} g_i^{(t)} g_{j|i}^{(t)}}{\sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P_{ij}^{(t)}}
\end{aligned}$$

   Similarly, we have that

$$\begin{aligned}
\mathbb{E}[z_i^{(t)}|\mathcal{X}] &= P(z_i^{(t)} = 1|y^{(t)}, x^{(t)}, \theta) \\
&= \frac{P(y^{(t)}|z_i^{(t)} = 1, x^{(t)}, \theta)P(z_i^{(t)} = 1|x^{(t)}, \theta)}{P(y^{(t)}|x^{(t)}, \theta)} \\
&= \frac{(\sum_j g_{j|i}^{(t)} P_{ij}^{(t)})g_i^{(t)}}{\sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P_{ij}^{(t)}}
\end{aligned}$$

   . Note that when finding the expected value of $z_{j|i}^{(t)}$, we need to only focus on its relationship with other networks connected to node i. Therefore we can have

$$\begin{aligned}
\mathbb{E}[z_{j|i}^{(t)}|\mathcal{X}] &= P(z_{j|i}^{(t)} = 1|y^{(t)}, x^{(t)}, \theta) \\
&= \frac{P_{ij}^{(t)} g_{j|i}^{(t)}}{\sum_j g_{j|i}^{(t)} P_{ij}^{(t)}}
\end{aligned}$$

2. Observing the expressions above, we know that $\mathbb{E}[z_i^{(t)}|\mathcal{X}] = h_i^{(t)}$, $\mathbb{E}[z_{j|i}^{(t)}|\mathcal{X}] = h_{j|i}^{(t)}$, and $\mathbb{E}[z_{ij}^{(t)}|\mathcal{X}] = h_i^{(t)} h_{j|i}^{(t)}$

(b)

We first take the natural logarithm of the likelihood over the entire dataset and obtain the following log likelihood

$$\ell(Y, Z|X, \theta) = \sum_t \sum_i \sum_j z_{ij}^{(t)} \log(g_i g_{j|i} P_{ij}(y^{(t)}|x^{(t)}, \theta_E))$$

Take the expected value over the above expression, from (a) we know that the expected log likelihood can be written as, using the notation from the paper,

$$Q(\theta, \theta^{(p)}) = \mathbb{E}(\ell(Y, Z|X, \theta^{(p)})) = \sum_t \sum_i \sum_j h_{ij}^{(t)} \log(g_i g_{j|i} P_{ij}^{(t)})$$

, where $\theta^{(p)}$ is the parameter values at the p-th iteration.

1. We first consider any specific U_{ij}.

   Note that the natural logarithm transforms the term

   $$\log(g_i g_{j|i} P_{ij}^{(t)}) = \log g_i + \log g_{j|i} + \log P_{ij}^{(t)}$$

   and the first two terms cannot be changed by U_{ij}.

   The term is then only related to the quantity

   $$\sum_t h_{ij}^{(t)} \log P_{ij}^{(t)}$$

   , so the problem for U_{ij} that we need to solve becomes

   $$U_{ij}^{(p+1)} = \arg\max_{U_{ij}} \sum_t h_{ij}^{(t)} \log P_{ij}^{(t)}$$

   . Another parameter that cannot affect the values of $g_i, g_{j|i}$ is σ_i, and the optimization problem for this term is

   $$\sigma_i^{(p+1)} = \arg\max_{\sigma_i} \sum_t \sum_j h_{ij}^{(t)} \log P_{ij}^{(t)}$$

   and the optimization problem for σ becomes

   $$\sigma^{(p+1)} = \arg\max_{\sigma} \sum_t \sum_i \sum_j h_{ij}^{(t)} \log P_{ij}^{(t)}$$

2. Expand out $\log P_{ij}^{(t)}$ first using the expression we have derived in part 3.1, we have

   $$\log P_{ij}^{(t)} = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma^2 + \sigma_i^2) - \frac{(y^{(t)} - U_{ij} X^{(t)})^T (y^{(t)} - U_{ij} X^{(t)})}{2(\sigma_i^2 + \sigma^2)}$$

3.