Homework7

Q1: Linear Dynamic System (Kalman Filter)

Derive the equations to learn the transition matrix F and observation matrix H

Based on the quote from the equation on lecture slide "OtherDirectedPGMFall20V6", page 74 "...state is not hidden during learning...", we know that in this Q1, all variables are observable.

Given state transition and observation matrix from the homework instruction, Q and R is known:

$$x^{t+1} = F * x^t + w, \quad w \sim N(0, Q)$$
$$y^t = H * x^t + v, \quad v \sim N(0, R)$$

This means we also know the prior distribution of state x_0.

$$p(x_0) = N(0, Q)$$

And meanwhile, we know the data we want to learn from $S = \{X^t, Y^t\}_{t=1}^T$.

Now we can define the objective function for learning problem of this LDS system as:

$$\begin{aligned} &\max_{F,H} LL(S; F, H) \\ =&\max_{F,H} \log p(X^{1:T}, Y^{1:T} | F, H) \end{aligned} \tag{1.1}$$

We see the LDS as a BN that is unrolled over time. To calculate the joint distribution, we can factorize the BN.

$$
\begin{aligned}
&\max_{F,H} \log p(X^{1:T}, Y^{1:T}|F,H) \\
&\text{Apply the state and observation independent assumption} \\
&= \max_{F,H} \log p_0(X^1, Y^1) \prod_{t=2}^{T} p(X^t, Y^t) \\
&= \max_{F,H} \log p_0(X^1) p(Y^1|X^1, H) \prod_{t=2}^{T} p(X^t|X^{t-1}, F) p(Y^t|X^t, H) \\
&= \max_{F,H} \log p_0(X^1) + \sum_{t=2}^{T} \log p(X^t|X^{t-1}, F) + \sum_{t=1}^{T} \log p(Y^t|X^t, H)
\end{aligned} \tag{1.2}
$$

Learning F and H is decomposable, because the whole sequence is observable. So the parameter learning of transition function can be solved by:

$$
\max_{F} \sum_{2}^{T} \log p(X^t|X^{t-1}, F) \tag{1.3}
$$

And parameter learning of observation matrix can be solved by:

$$
\max_{H} \sum_{1}^{T} \log p(Y^t|X^t, H) \tag{1.4}
$$

Equation 1.3 and equation 1.4 are the key equations in parameter learning of LDS. We can find the close-form solution both key equations by taking the advantage of Gaussian assumption of both distribution.

1.1 Close-form solution to parameter (F) learning of transition function

Apply the Gaussian assumption, equation 1.3 can be expand as:

$$
\begin{aligned}
&\max_F \sum_2^T \log p(X^t|X^{t-1}, F) \\
=&\max_F \sum_2^T \log N(X^t; F * X^{t-1}, Q) \\
\Rightarrow &\max_F \sum_2^T \log \exp -\frac{(X^t - F * X^{t-1})^\top Q^{-1}(X^t - F * X^{t-1})}{2} \\
\Rightarrow &\min_F \sum_{t=2}^T (X^t - F * X^{t-1})^\top (X^t - F * X^{t-1}) \\
=&\min_F \sum_{t=2}^T {X^t}^\top X^t - 2F \sum_{t=2}^T {X^t}^\top X^{t-1} + F^\top F \sum_{t=2}^T {X^{t-1}}^\top X^{t-1} \\
=&\min_F f_1(F)
\end{aligned} \tag{1.5}
$$

When the derivative of $f_1(F) = 0$, equation 1.5 reach the minimum.

$$
\begin{aligned}
&f_1'(F) \\
=&2F \sum_{t=2}^T {X^{t-1}}^\top X^{t-1} - 2 \sum_{t=2}^T {X^t}^\top X^{t-1} \\
\Rightarrow &F^* = \frac{\sum_{t=2}^T {X^{t-1}}^\top X^{t-1}}{\sum_{t=2}^T {X^t}^\top X^{t-1}} \\
\Rightarrow &F^* = \sum_{t=2}^T {X^t}^\top X^{t-1} [\sum_{t=2}^T {X^{t-1}}^\top X^{t-1}]^{-1}
\end{aligned}
$$

The close-form solution to equation 1.3 is $F^* = \sum_{t=2}^T {X^t}^\top X^{t-1} [\sum_{t=2}^T {X^{t-1}}^\top X^{t-1}]^{-1}$

1.2 Close-form solution to parameter (H) learning of observation function

Solving equation 1.4 is very similar to solving equation 1.3.

$$
\begin{aligned}
&\max_H \sum_{t=1}^{T} \log p(Y^t|X^t, H) \\
=&\max_F \sum_{1}^{T} \log N(Y^t; H * X^t, R) \\
\Rightarrow &\max_H \sum_{1}^{T} \log \exp -\frac{(Y^t - H * X^t)^\top R^{-1} (Y^t - H * X^t)}{2} \\
\Rightarrow &\min_H \sum_{t=1}^{T} (Y^t - H * X^t)^\top (Y^t - H * X^t) \\
=&\min_H \sum_{t=1}^{T} {Y^t}^\top Y^t - 2H \sum_{t=1}^{T} {Y^t}^\top X^t + H^\top H \sum_{t=1}^{T} {X^t}^\top X^t \\
=&\min_H f_2(H)
\end{aligned}
\tag{1.6}
$$

Similarly, we can get the close from solution to equation 1.6 by letting $f'2(H) = 0$:

$$
H^* = \sum_{t=1}^{T} {Y^t}^\top X^t [\sum_{t=1}^{T} {X^t}^\top X^t]^{-1}
$$

Q2. Influence Diagrams

Evidence e: $X_2 = 1$

According to the equation on lecture slide “OtherDirectedPGMFall20V6”, page 86, we know that the objective function of ID is to maximize the expected utility:

$$
\arg\max_d EU(d|e) = \arg\max_d \sum_x p(x|d, e) U(x, d) \tag{2.1}
$$

When given the PGM structure, we can expand the term $p(x|d, e)$ as:

$$
\begin{aligned}
p(x|d,e) =& p(x_1, x_3|d, x_2 = 1) \\
=& \frac{p(x_1, x_3, d, x_2 = 1)}{p(d, e = 2)} \\
=& \frac{p(x_1, x_3, d, x_2 = 1)}{\sum_{x_1, x_3} p(x_1, x_3, d, x_2 = 1)} \\
=& \frac{p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)}{\sum_{x_1, x_3} p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)} \qquad \text{Apply the } I(\mathcal{G})
\end{aligned}
$$

and $U(x, d) = U(x_3, d)$. In this case, we can specify the objective function in this Q2 as:

$$
\begin{aligned}
& \arg\max_d EU(d|x_2 = 1) \\
&= \sum_{x_1, x_3} \frac{p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)}{\sum_{x_1, x_3} p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)} U(x_3, d) \qquad (2.2) \\
&= \sum_{x_1, x_3} f_1(x_1, x_3, d) U(x_3, d)
\end{aligned}
$$

Where $p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)$:

	$X_3 = 0$	$X_3 = 1$
$X_1 = 0, D = 0$	.1568	.0392
$X_1 = 1, D = 0$	.0144	.0336
$X_1 = 0, D = 1$	.0672	.0168
$X_1 = 1, D = 1$	.0216	.0504

Where $p(d, x_2 = 1) = \sum_{x_1, x_3} p(x_1)p(x_2 = 1|x_1)p(x_3|x_1)p(d|x_1)$ is:

$X_3 = 0$	$X_3 = 1$
.244	.156

And $f_1(x_1, x_3, d) = p(x_1, x_3|D)$:

	$X_3 = 0$	$X_3 = 1$
$X_1 = 0, D = 0$	.642623	.160656

$X_1 = 1, D = 0$	.059016	.137705
$X_1 = 0, D = 1$	.430769	.107692
$X_1 = 1, D = 1$	.138462	.323077

Then $EU(d\|x_2 = 1)$**:**	
$D = 0$	$D = 1$
55.967213	65.846154

Given this $EU(d|x_2 = 1)$, it is easy to conclude that the decision that return the maximized expected utility is $d = 1$.

Q3 Hierarchical Bayesian model

3.1 What types of distributions $p(x|\theta)$ and $p(\theta|\alpha)$ are?

Because we know that X nodes are all binary node $p(x|\theta)$ are Bernoulli distribution. $p(\theta|\alpha)$ are beta distributions.

3.2 Write down the joint distribution of the random variables (Xs) and random parameters (θ s)as a function of the hyper-parameters (α s)

Joint distribution of random variables and random parameters can be written as:

$$f(\vec{\alpha}) = p(\vec{X}, \vec{\theta}|\vec{\alpha}) = p(X_1, X_2, X_3, \theta_1, \theta_2, \theta_3|\alpha_1, \alpha_2, \alpha_3) \tag{3.1}$$

Taking the advantage of the independent properties $I(\mathcal{G})$ of the given structure, we can factorize the joint distribution equation 3.1 using the chain rule (follow the optimal sequence):

$$\begin{aligned} & f(\vec{\alpha}) \\ =& p(\theta_1|\alpha_1)p(\theta_2|\alpha_2)p(\theta_3|\alpha_3) \\ & p(X_2|\theta_2)p(X_1|\theta_1, X_2)p(X_3|\theta_3, X_2, X_1) \end{aligned} \tag{3.2}$$

Q4 Latent Dirichlet Allocation (LDA)

4.1 Given training data $D = w_i, y_i$, where $i = 1, 2.., N$ documents and wi are the words in document, i and $w_i = w_{ij}$, where $j = 1, 2, ..., M$ words. Provide the key equations to solve for the hyper parameters α, β, and γ.

According to the equation on lecture slide "OtherDirectedPGMFall20V6", page 136, we can list the following properties of the distributions:

- $\theta_i \sim Dir(\alpha)$
- $z_{ij} \sim Multi(z_{ij}|\theta_i)$
- $w_{ij} \sim Multi(w_{ij}|z_{ij})$

Adding the make-up material professor added in the class, we know

- $y_i \sim Dir(\gamma)$

To solve hyper-parameters α and γ and parameter β is a learning problem. The key of solving the learning problem is to find the log likelihood function of the data.

$$\arg\max_{\alpha,\beta,\gamma} \log p(D|\alpha, \beta, \gamma) \tag{4.1}$$

Where $\log p(D|\alpha, \beta, \gamma)$ can be factorized as:

$$
\begin{aligned}
&\max_{\alpha,\beta,\gamma} \log p(D|\alpha,\beta,\gamma) \\
=&\max_{\alpha,\beta,\gamma} \log \prod_{i=1}^{N} p(w_i, y_i|\alpha,\beta,\gamma) \\
=&\max_{\alpha,\beta,\gamma} \log \prod_{i=1}^{N}\prod_{j=1}^{M} p(w_{ij}, y_i|\alpha,\beta,\gamma) \\
=&\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N}\sum_{j=1}^{M} \log \int_{\theta_i} \sum_{z_{ij}} p(w_{ij}, y_i, \theta_i, z_{ij}|\alpha,\beta,\gamma) d_{\theta_i} \\
=&\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N}\sum_{j=1}^{M} \log \int_{\theta_i} \sum_{z_{ij}} q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \qquad (4.2) \\
&\frac{p(w_{ij}, y_i, \theta_i, z_{ij}|\alpha,\beta,\gamma)}{q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1})} d_{\theta_i} \\
\geq&\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N}\sum_{j=1}^{M} \int_{\theta_i} \sum_{z_{ij}} q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \\
&\log \frac{p(w_{ij}, y_i, \theta_i, z_{ij}|\alpha,\beta,\gamma)}{q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1})} d_{\theta_i} \quad \text{Jensen's inequality} \\
\Rightarrow&\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N}\sum_{j=1}^{M} \int_{\theta_i} \sum_{z_{ij}} q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \log p(w_{ij}, y_i, \theta_i, z_{ij}|\alpha,\beta,\gamma) d_{\theta_i}
\end{aligned}
$$

What we do in equation 4.2 is to apply Expected-Maximization method, to infer the latent variables θ_i and z_{ij}, and to raise the lower bound of the log-likelihood iteratively. Where the E-step is:

$$
\begin{aligned}
&\log Q(\alpha,\beta,\gamma|\alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \\
=&\sum_{i=1}^{N}\sum_{j=1}^{M} \int_{\theta_i} \sum_{z_{ij}} q(\theta_i, z_{ij}|w_{ij}, y_i, \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \\
&\log p(w_{ij}, y_i, \theta_i, z_{ij}|\alpha,\beta,\gamma) d_{\theta_i} \qquad (4.3) \\
=&\sum_{i=1}^{N}\sum_{j=1}^{M} \int_{\theta_i} q(\theta_i|y_i, \alpha^{t-1}, \gamma^{t-1}) \sum_{z_{ij}} q(z_{ij}|, w_{ij}, \theta_i, \beta^{t-1}) \\
&\log p(y_i|\gamma) p(\theta_i|\alpha, y_i) p(z_{ij}|\theta_i) p(w_{ij}|z_{ij}, \beta) d_{\theta_i}
\end{aligned}
$$

And the M-step is:

$$\alpha^t, \beta^t, \gamma^t = \arg\max_{\alpha,\beta,\gamma} Q(\alpha, \beta, \gamma | \alpha^{t-1}, \beta^{t-1}, \gamma^{t-1}) \tag{4.4}$$

We need to do both E-step and M-step iteratively until convergence of the learned parameters.

For equation 4.4, we can hardly find the analytical closed-form solution. To solve it, we can either use sampling method (Gibbs sampling or HM sampling), or variational method to approximate them. Another method is to use Laplace approximation, and this may turn the EM algorithm to a hard-EM algorithm.

4.2 Given the words w_i' for document i', provide the key questions to infer the document's label y_i.

- The evidence is $w_{ij} = w'_{ij}$
- The query variable is y_i.
- The non-query unknown variables are $\alpha, \beta, \gamma, \theta_i, z_{ij}$

The inference in Hierarchical Bayesian model includes fully Bayesian inference and empirical Bayesian inference

4.2.1 Fully Bayesian inference

The objective function is:

$$
\begin{aligned}
& p(y_i|w_i = w_i') \\
= & \int_{\alpha,\beta,\gamma,\theta_i} \sum_{z_i} p(y_i, \alpha, \beta, \gamma, \theta_i, z_i | w_i = w_i') d_\alpha d_\beta d_\gamma d_{\theta_i} \\
= & \int_{\alpha,\beta,\gamma,\theta_i} \sum_{z_{ij}} \prod_{j=1}^{M} p(y_i, \alpha, \beta, \gamma, \theta_i, z_{ij} | w_{ij} = w_{ij}') d_\alpha d_\beta d_\gamma d_{\theta_i} \\
\propto & \int_{\alpha,\beta,\gamma,\theta_i} \sum_{z_{ij}} \prod_{j=1}^{M} p(y_i, \alpha, \beta, \gamma, \theta_i, z_{ij}, w_{ij} = w_{ij}') d_\alpha d_\beta d_\gamma d_{\theta_i} \\
= & \int_{\gamma} p(\gamma) p(y_i|\gamma) \int_{\alpha} p(\alpha) \int_{\theta_i} p(\theta_i|\alpha, y_i) \sum_{z_{ij}} \prod_{j=1}^{M} p(z_{ij}|\theta_i) \\
& \int_{\beta} p(\beta) p(w_{ij} = w_{ij}'|z_{ij}, \beta) d_\alpha d_\beta d_\gamma d_{\theta_i}
\end{aligned}
\tag{4.5}
$$

Note that, according to “OtherDirectedPGMFall20V6”, page 136, the distribution of hyperparameter distribution is dirichlet and distribution for topics and words are all multinomial distribution, which means the hyperparameters are conjugate priors. This conclusion allows us to calculate $p(y_i|w_i = w_i')$ without calculating the distribution of evidence.

4.2.1 Empirical Bayesian inference

In question 4.1, we have learned the optimal hyperparameter $\alpha^*, \beta^*, \gamma^*$ in the training set. Here, we can use these learned optimal hyperparameter to infer y_i given w_i'

$$
\begin{aligned}
&p(y_i|w_i = w_i', \alpha^*, \beta^*, \gamma^*) \\
=&\int_{\theta_i} \sum_{z_i} p(y_i, \theta_i, z_i|w_i = w_i', \alpha^*, \beta^*, \gamma^*) d_{\theta_i} \\
=&\int_{\theta_i} \sum_{z_{ij}} \prod_{j=1}^{M} p(y_i, \theta_i, z_{ij}|w_{ij} = w_{ij}', \alpha^*, \beta^*, \gamma^*) d_{\theta_i} \\
\propto&\int_{\theta_i} \sum_{z_{ij}} \prod_{j=1}^{M} p(y_i, \theta_i, z_{ij}, w_{ij} = w_{ij}'|\alpha^*, \beta^*, \gamma^*) d_{\theta_i} \\
=&p(y_i|\gamma^*) \int_{\theta_i} p(\theta_i|\alpha^*, y_i) \sum_{z_{ij}} \prod_{j=1}^{M} p(z_{ij}|\theta_i) \\
&p(w_{ij} = w_{ij}'|z_{ij}, \beta^*) d_{\theta_i}
\end{aligned}
\tag{4.6}
$$

Similar to 4.2.1 we can take the advantage of conjugate property of LDA to simplify the calculation.