NATIONAL UNIVERSITY OF SINGAPORE

CS5340 - Uncertainty Modeling in AI

(Quiz 2, Semester 2 AY2021/22)

SOLUTIONS

Time Allowed: 1 hour

Instructions

- This is an open-book quiz. You may refer to any of the lecture slides and tutorials.
- You may *not* refer to any external online material or use any software to help you answer the questions.
- Please do not cheat; your answers *must* be your own. Do *not* collaborate with anyone else.
- Please put all your answers in Luminus.
- Read each question *carefully*. Don't get stuck on any one problem. The questions are *not* in any particular order of difficulty.
- Don't panic. The problems often look more difficult than they really are.
- Good luck!

Student Number.: ________________________________

Common Probability Distributions

Distribution (Parameters)	PDF/PMF
Normal (μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
Bernoulli (r)	$r^x(1-r)^{(1-x)}$
Categorical (π)	$\prod_{k=1}^K \pi_k^{x_k}$
Binomial (μ, N)	$\binom{N}{x}\mu^x(1-\mu)^{N-x}$
Poisson (λ)	$\frac{\lambda^x \exp[-\lambda]}{x!}$
Beta (α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Gamma (a, b)	$\frac{1}{\Gamma(a)} b^a x^{a-1} \exp[-bx]$
Dirichlet $(\boldsymbol{\alpha})$	$\frac{\Gamma(\sum_k^K \alpha_k)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^K x_k^{\alpha_k-1}$
Multivariate Normal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\frac{1}{(2\pi)^{D/2}\|\Sigma\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$
Uniform (a, b)	$\frac{1}{b-a}$
Cauchy (x_0, γ)	$\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$

Note: $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the Gamma function.

1 True or False?

For the following questions, please answer TRUE or FALSE.

Problem 1. [1 points] Let $Z = aX + bY$ where $X \sim \text{Bernoulli}(0.2)$ and $Y \sim \mathcal{N}(2, 1)$. Then

$$\mathbb{E}[Z] = \frac{a}{5} + 2b$$

Solution: True.

Problem 2. [1 points] Let the function $f(X) = -(X^2)$ and $x \sim p(X)$. Define a new distribution $q(X)$ with the same support as $p(X)$. Then,

$$\mathbb{E}_p[f(X)] = \mathbb{E}_q[f(X)p(X)/q(X)]$$

Solution: True

Problem 3. [1 points] Consider $x \sim \text{Beta}(\alpha, \beta)$. Binomial prior distributions over α and β are conjugate to a Beta likelihood and would lead to tractable and closed-form Bayesian inference. In particular, the new parameters are computed as $\alpha' = \alpha + s$ and $\beta' = \beta + (n - s)$ where s is the number of 1's observed and n is the number of samples.

Solution: False. α and β should be continuous, but Binomial prior only support discrete values.

Problem 4. [1 points] The Markov blanket for a node in a Markov Random Field is the set containing its neighbors.

Solution: True.

Problem 5. [2 points] For any independent variables X and Y,

$$\mathbb{E}[X^2 + Y^2] = \mathbb{V}[X] + \mathbb{V}[Y] + (\mathbb{E}[X] + \mathbb{E}[Y])^2 - 2\mathbb{E}[X]\mathbb{E}[Y]$$

Solution: True.

Problem 6. [1 points] The Poisson distribution is in Exponential Family.

Solution: True.

Problem 7. [1 points] An Exponential Family distribution always has a conjugate prior.

Solution: True.

Problem 8. [2 points] True or False:

$$(X \perp Y|Z, W) \Rightarrow (X \perp Y|Z) \wedge (X \perp Y|W)$$

In other words, $(X \perp Y|Z, W)$ implies $(X \perp Y|Z)$ and $(X \perp Y|W)$.

Solution: False.

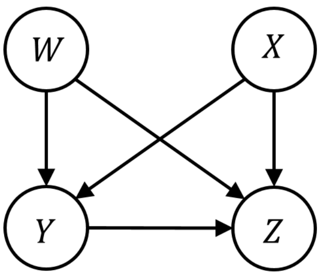
Problem 9. [2 points] True or False:

$$(X \perp Y|Z, W) \wedge (X \perp Y|W) \Rightarrow (X \perp Y|Z)$$

In other words, if $(X \perp Y|Z, W)$ and $(X \perp Y|W)$ then $(X \perp Y|Z)$.

Solution: False.

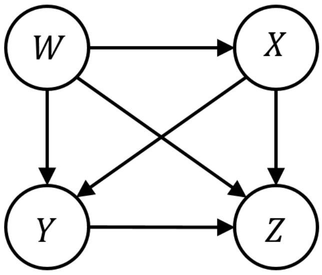
Problem 10. [2 points] Consider a Bayesian Network with four nodes W, X, Y, and Z and the following facts:



- $W \sim \text{Normal}(2, \sigma^2)$
- $X \sim \text{Normal}(0, v^2)$
- $Y = aW + bX$ where a and b are scalars.
- $Z = cW + dX + eY$ where c, d and e are scalars.

Solution: True. Then, the conditional $p(Y|W, X, Z)$ is Gaussian, since this is a linear Gaussian model that we have covered in the tutorial.

Problem 11. [2 points] Consider a Bayesian Network with three nodes W, X, Y, and Z and the following facts:



- $W \sim \text{Normal}(2, \sigma^2)$
- $X = (\mathbf{u}^\top f_\theta(\mathbf{k})) + W$ where f_θ is a neural network. $\mathbf{u}$ and $\mathbf{k}$ are real vectors $\mathbf{u}, \mathbf{k} \in \mathbb{R}^d$.
- $Y = aW + bX$ where a and b are scalars.
- $Z = W + X + Y$

Solution: True. Then, the conditional $p(Y|W, X, Z)$ is Gaussian, since this is a linear Gaussian model that we have covered in the tutorial.

2 Valid Transitions

For each of the matrices below, select True if the matrix is ergodic. Select False otherwise.

Problem 12. [1 points]

$$T = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.9 & 0.1 & 0.0 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}$$

Solution: Yes, this transition matrix is aperiodic and irreducible.

Problem 13. [1 points]

$$T = \begin{bmatrix} 0.0 & 0.7 & 0.3 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.6 & 0.2 \end{bmatrix}$$

Solution: Yes, the transition matrix is aperiodic and irreducible.

Problem 14. [1 points]

$$T = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 \\ 0.0 & 0.3 & 0.7 \end{bmatrix}$$

Solution: No, the matrix is not irreducible.

Problem 15. [1 points]

$$T = \begin{bmatrix} 0.0 & 0.3 & 0.7 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

Solution: No, the matrix is not aperiodic.

Problem 16. [1 points]

$$T = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.5 & 0.0 & 0.5 \\ 0.0 & 1.0 & 0.0 \end{bmatrix}$$

Solution: No, the matrix is not aperiodic.

3 More MCMC

Instead of specifying the proposal distribution, we can specify a "proposal transition function" for a MCMC sampler. In other words, we transform x to a new candidate sample x' via some function.

Problem 17. [3 points] Which of the following functions will lead to sampling from the stationary distribution given properly set hyperparameters? The hyperparameters are assumed *constant* throughout the chain. Assume the target distribution to be a **continuous** univariate distribution. Select all that apply.

A. $x' = x + \epsilon$ where $\epsilon \sim \text{Normal}(0, \sigma^2)$
B. $x' = x + (-1)^s \epsilon$ where $\epsilon \sim \text{Beta}(a, b)$ and $s \sim \text{Bernoulli}(0.5)$
C. $x' = x + \epsilon$ where $\epsilon \sim \text{Poisson}(a)$
D. $x' = x + (-1)^s \epsilon$ where $\epsilon \sim \text{Bernoulli}(r)$ and $s \sim \text{Bernoulli}(0.5)$
E. $x' = x + (-1)^s \epsilon$ where $\epsilon \sim \text{Gamma}(a, b)$ and $s \sim \text{Bernoulli}(0.5)$

Solution: A, B and E are valid proposal transition functions. C is not irreducible, since $\epsilon \sim \text{Poisson}(a)$ will always be larger than 0, therefore, it can not reach the value that is smaller than the x. D is also not irreducible, since Bernoulli is discrete, therefore if we start at a state x, we will never be able to reach the all the real numbers in the future, for example $x + 0.1$.

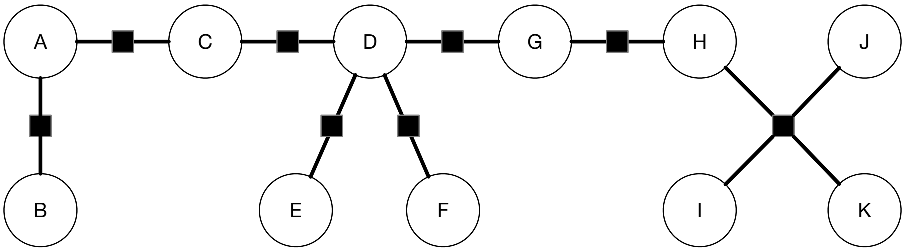
Problem 18. [3 points] Which of the following functions will lead to sampling from the stationary distribution given properly set hyperparameters? The hyperparameters are assumed *constant* throughout the chain. Assume the target distribution to be a **continuous** univariate distribution with strictly **positive support**, i.e., $p(x \leq 0) = 0$. Select all that apply.

A. $x' = x + \epsilon$ where $\epsilon \sim \text{Normal}(0, \sigma^2)$
B. $x' = x + (-1)^s \epsilon$ where $\epsilon \sim \text{Beta}(a, b)$ and $s \sim \text{Bernoulli}(0.5)$
C. $x' = x + \epsilon$ where $\epsilon \sim \text{Poisson}(a)$
D. $x' = x + (-1)^s \epsilon$ where $\epsilon \sim \text{Bernoulli}(r)$ and $s \sim \text{Bernoulli}(0.5)$
E. $x' = x + \epsilon$ where $\epsilon \sim \text{Gamma}(a, b)$

Solution: A, B. C and D is not irreducible. E is not irreducible, since $\epsilon \sim \text{Gamma}(a, b)$ will always be larger than 0, therefore, we will never be able to reach the number that is smaller than x.

4 Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 19. [1 points] Sample A.

A. $p(A)$ (sample from the prior)

B. $p(A|B, C)$

C. $p(A|B, C, D)$

D. $p(A|B, C, D, E)$

E. $p(A|B)$

F. None of the other answers is correct.

Solution: $p(A|B, C)$

Problem 20. [1 points] Sample D.

A. $p(D|C, G, E, F)$

B. $p(D|A, C, G, E, F, H)$

C. $p(D|A, C, D)$

D. $p(D)$

E. $p(D|E, F)$

F. None of the other answers is correct.

Solution: $p(D|C, E, F, G)$

Problem 21. [1 points] Sample E.

A. $p(E)$

B. $p(E|D)$

C. $p(E|C, D, F, G)$

D. $p(E|D, F)$

E. $p(E|A, B, C, D)$

F. None of the other answers is correct.

Solution: $p(E|D)$

Problem 22. [1 points] Sample H.

A. $p(H)$

B. $p(H|D, G)$

C. $p(H|G, I, J, K)$

D. $p(H|G)$

E. $p(H|I, J, K)$

F. None of the other answers is correct.

Solution: $p(H|G, I, J, K)$

Problem 23. [1 points] Sample K.

A. $p(K)$

B. $p(K|H, I, J)$

C. $p(K|H)$

D. $p(K|G, H, I, J, K)$

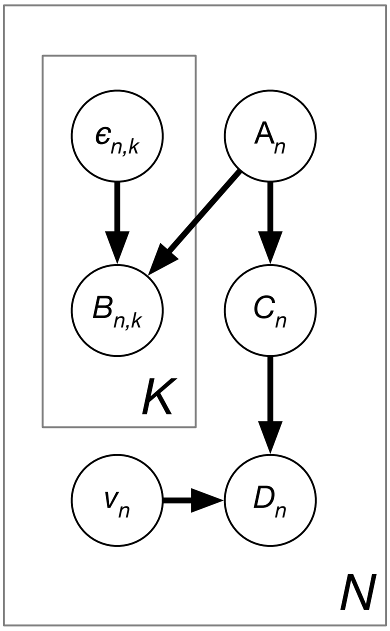
E. $p(K|A, C, D, F, H)$

F. None of the other answers is correct.

Solution: $p(K|H, I, J)$

5 A Regression Model

Consider the following DGM,



along with the following distributions and relationships between variables:

- $A_n \sim \text{Normal}(0, 1)$
- $\epsilon_{n,k} \sim \text{Normal}(0, \sigma_\epsilon^2)$
- $v_n \sim \text{Normal}(0, \sigma_v^2)$
- $B_{n,k} = wA_n + \epsilon_{n,k}$ where w is a scalar.
- $C_n = rA_n + \mu$ where r, μ are scalars.
- $D_n = C_n + v_n$

For convenience, let us define the following:

- The parameters of this model is the set $\theta = \{w, r, \mu, \sigma_\epsilon^2, \sigma_v^2\}$.
- Let $B_n = \{B_{n,k}\}_{k=1}^K$, i.e., for a given n, $B_n = \{B_{n,1}, B_{n,2}, \ldots B_{n,K}\}$
- Likewise, let $\epsilon_n = \{\epsilon_{n,k}\}_{k=1}^K$
- $\mathcal{X}$ is the set of the random variables $\mathcal{X} = \{A_n, B_n, \epsilon_n, C_n, v_n, D_n\}_{n=1}^N$

This is a linear regression model extended such that we may not observe A_n but instead only noisy observations B_n of it. Likewise, the targets D_n are corrupted by noise.

For the questions in this section, assume that θ are deterministic parameters (not random variables) and θ is known.

Problem 24. [2 points] Which of the following joint distributions corresponds to the given model?

A. $p(\mathcal{X}) = \prod_n^N p(D_n|C_n, v_n)p(C_n|A_n)p(A_n)p(v_n) \prod_k^K p(B_{n,k}|A_n, \epsilon_{n,k})p(\epsilon_{n,k})$

B. $p(\mathcal{X}) = \prod_n^N p(A_n, C_n|D_n, v_n)p(D_n)p(v_n) \prod_k^K p(B_{n,k}|D_n, C_n, \epsilon_{n,k})p(\epsilon_{n,k})$

C. $p(\mathcal{X}) = \prod_n^N p(D_n, C_n|A_n, v_n)p(A_n)p(v_n) \prod_k^K p(A_n|B_{n,k}, \epsilon_{n,k})p(\epsilon_{n,k})$

D. $p(\mathcal{X}) = \prod_n^N p(D_n|A_n, v_n)p(A_n)p(v_n)$

E. $p(\mathcal{X}) = p(D_n|A_n, v_n)p(A_n)p(v_n)p(B_{n,k}|A_n, \epsilon_{n,k})p(\epsilon_{n,k})$

F. None of the other answers is correct.

Solution: A. $p(\mathcal{X}) = \prod_n^N p(D_n|C_n, v_n)p(C_n|A_n)p(A_n)p(v_n) \prod_k^K p(B_{n,k}|A_n, \epsilon_{n,k})p(\epsilon_{n,k})$

Problem 25. [2 points] What is the covariance of A_i and $B_{j,k}$ where $i \neq j$?

A. 0

B. 1

C. w

D. r

E. wr

F. w^2r^2

G. None of the other answers is correct.

Solution: 0. From the graph, A_i and $B_{j,k}$ are independent if $i \neq j$. Therefore, $\mathrm{Cov}(A_i, B_{j,k}) = 0$.

Problem 26. [2 points] What is the covariance of A_n and $B_{n,k}$ for a given n and k?

A. 0

B. 1

C. w

D. r

E. wr

F. w^2r^2

G. None of the other answers is correct.

Solution: $\mathrm{Cov}(A_n, B_{n,k}) = w$

$$\begin{aligned}
\mathrm{Cov}(A_n, B_{n,k}) &= \mathbb{E}[A_n - \mathbb{E}[A_n]][B_{n,k} - \mathbb{E}[B_{n,k}]] && (1)\\
&= \mathbb{E}[A_n \cdot [wA_n + \epsilon_{n,k} - \mathbb{E}[wA_n + \epsilon_{n,k}]] && (2)\\
&= \mathbb{E}[wA_n^2 + A_n\epsilon_{n,k}] && (3)\\
&= w\mathbb{E}[A_n^2] + \mathbb{E}[A_n\epsilon_{n,k}] && (4)\\
&= w\mathbb{E}[A_n^2] && (5)
\end{aligned}$$

Since $\mathrm{Var}[A_n] = -\mathbb{E}[A_n]^2 + \mathbb{E}[A_n^2] = 1$, so $\mathbb{E}[A_n^2] = 1$, and $\mathrm{Cov}(A_n, B_{n,k}) = w$.

Problem 27. [3 points] What is the variance of $B_{n,k}$?

A. 0

B. 1

C. $\sigma_\epsilon^2 + w^2$

D. $\sigma_v^2 + r^2$

E. $\sigma_\epsilon^2 + r^2$

F. $\sigma_v^2 + w^2$

G. None of the other answers is correct.

Solution: $\sigma_\epsilon^2 + w^2$

$$\mathrm{Var}[B_{n,k}] = \mathbb{E}[B_{n,k}^2] - \mathbb{E}[B_{n,k}]^2$$
$$\mathbb{E}[B_{n,k}]^2 = \mathbb{E}[wA_n + \epsilon_{n,k}]^2 = 0$$

$$\begin{aligned}
\mathbb{E}[B_{n,k}^2] &= \mathbb{E}[w^2 A_n^2 + 2wA_n\epsilon_{n,k} + \epsilon_{n,k}^2] && (6)\\
&= w^2\mathbb{E}[A_n^2] + 2w\mathbb{E}[A_n\epsilon_{n,k}] + \mathbb{E}[\epsilon_{n,k}^2] && (7)\\
&= w^2 + \sigma_\epsilon^2 && (8)
\end{aligned}$$

Since $\mathrm{Var}[\epsilon_{n,k}] = \mathbb{E}[\epsilon_{n,k}^2] - \mathbb{E}[\epsilon_{n,k}]^2 = \sigma_\epsilon^2$, $\mathbb{E}[\epsilon_{n,k}^2] = \sigma_\epsilon^2$.

Problem 28. [3 points] What is the covariance of $B_{n,k}$ and D_n?

A. 0

B. 1

C. wr

D. w^2r^2

E. $\sigma_\epsilon^2 + \sigma_v^2$

F. $w\sigma_\epsilon^2 + r\sigma_v^2$

G. None of the other answers is correct.

Solution: wr
Since $D_n = rA_n + \mu + v_n$ and $B_{n,k} = wA_n + \epsilon_{n,k}$,

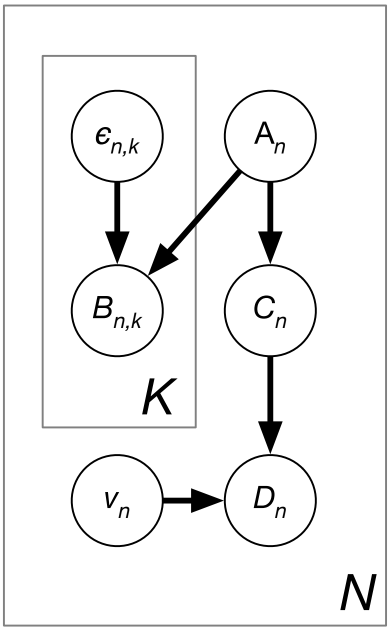
$$\begin{aligned}
\mathrm{Cov}(B_{n,k}, D_n) &= \mathbb{E}[B_{n,k} - \mathbb{E}[B_{n,k}]][D_n - \mathbb{E}[D_n]] && (9) \\
&= \mathbb{E}[wA_n + \epsilon_{n,k}][rA_n + \mu + v_n - \mu] && (10) \\
&= wr\mathbb{E}[A_n^2] + w\mathbb{E}[A_n v_n] + r\mathbb{E}[\epsilon_{n,k} A_n] + \mathbb{E}[\epsilon_{n,k} v_n] && (11)
\end{aligned}$$

Since A_n and v_n are independent, $\mathbb{E}[A_n v_n] = \mathbb{E}[A_n]\mathbb{E}[v_n] = 0$,
Since A_n and $\epsilon_{n,k}$ are independent, $\mathbb{E}[\epsilon_{n,k} v_n] = \mathbb{E}[\epsilon_{n,k}]\mathbb{E}[A_n] = 0$,
Since $\epsilon_{n,k}$ and v_n are independent, $\mathbb{E}[\epsilon_{n,k} v_n] = \mathbb{E}[\epsilon_{n,k}]\mathbb{E}[v_n] = 0$.
As a result, $\mathrm{Var}(B_{n,k}, D_n) = wr\mathbb{E}[A_n^2] = wr$.

6 Learning Parameters

We will reuse the same model as in Section 5.

Consider the following DGM,



along with the following distributions and relationships between variables:

- $A_n \sim \text{Normal}(0, 1)$
- $\epsilon_{n,k} \sim \text{Normal}(0, \sigma_\epsilon^2)$
- $v_n \sim \text{Normal}(0, \sigma_v^2)$
- $B_{n,k} = wA_n + \epsilon_{n,k}$ where w is a scalar.
- $C_n = rA_n + \mu$ where r, μ are scalars.
- $D_n = C_n + v_n$

For convenience, let us define the following:

- The parameters of this model is the set $\theta = \{w, r, \mu, \sigma_\epsilon^2, \sigma_v^2\}$.
- Let $B_n = \{B_{n,k}\}_{k=1}^K$, i.e., for a given n, $B_n = \{B_{n,1}, B_{n,2}, \ldots B_{n,K}\}$
- Likewise, let $\epsilon_n = \{\epsilon_{n,k}\}_{k=1}^K$
- $\mathcal{X}$ is the set of the random variables $\mathcal{X} = \{A_n, B_n, \epsilon_n, C_n, v_n, D_n\}_{n=1}^N$

This is a linear regression model extended such that we may not observe A_n but instead only noisy observations B_n of it. Likewise, the targets D_n are corrupted by noise.

NOTE: For the questions in this section, θ is **unknown** and we wish to learn it from data. Assume that θ are deterministic parameters (not random variables).

Problem 29. [2 points] Suppose we observe A_n, C_n, D_n and we only want to learn the parameter r via MLE. Which of the following should we compute? If multiple solutions are similarly desirable, pick the set with the smallest number of random variables.

A. $\arg\max_r \sum_n \log p(C_n|A_n, r)$

B. $\arg\max_r \sum_n \log p(C_n, D_n|A_n, r)$

C. $\arg\max_r \sum_n \log p(A_n|r)$

D. $\arg\max_r \sum_n \log p(A_n, C_n, D_n|r)$

E. $\arg\max_r \sum_n \log p(C_n|r)$

F. None of the other answers is correct.

Solution: A. $\arg\max_r \sum_n \log p(C_n|A_n, r)$

Problem 30. [2 points] Suppose we only want to learn the parameter w via MLE. Which among the following random variables would we prefer to observe? If multiple solutions are similarly desirable, pick the set with the smallest number of random variables.

A. $\{A_n, B_n, C_n, D_n\}_{n=1}^N$

B. $\{B_n, C_n\}_{n=1}^N$

C. $\{A_n, B_n, \epsilon_n\}_{n=1}^N$

D. $\{B_n, C_n, \epsilon_n\}_{n=1}^N$

E. $\{B_n, C_n, D_n, \epsilon_n\}_{n=1}^N$

F. $\{A_n, B_n, C_n, D_n, \epsilon_n, v_n\}_{n=1}^N$

G. None of the other answers is correct.

Solution: C. $\{A_n, B_n, \epsilon_n\}_{n=1}^N$

Problem 31. [2 points] Suppose we only observe $\mathcal{O} = \{B_n, D_n\}_{n=1}^N$ and want to learn the parameters θ via MLE. Which of the following statements is correct?

A. The likelihood is tractable and we can directly optimize it using an off-the-shelf optimizer.

B. The likelihood is intractable due to the latent variables. We can learn the parameters via EM.

C. The likelihood is intractable due to the latent variables. Also, the posterior over the latent variables is intractable. We can learn the parameters via Monte-Carlo EM or variational inference.

D. None of the other statements is correct.

Solution: A. This is a variant of the linear Gaussian model. The likelihood is tractable and we can directly optimize it using an off-the-shelf optimizer.

Problem 32. [2 points] Suppose we only observe $\mathcal{O} = \{B_n, D_n\}_{n=1}^N$ and want to learn the parameters θ via Expectation Maximization. Consider that we first simplify the model by analytically marginalizing out ϵ_n, v_n, and C_n. Which of the following posteriors is needed to form the $Q(\theta, \theta^{old})$ function? Pick the best answer among the following.

A. $\prod_n^N p(A_n|B_n, D_n, \theta^{old})$

B. $\prod_n^N p(A_n, B_n, D_n, \theta^{old})$

C. $\prod_n^N p(C_n, \epsilon_n, v_n|B_n, D_n, \theta^{old})$

D. $\prod_n^N p(B_n, \epsilon_n, v_n|A_n, D_n, \theta^{old})$

E. $\prod_n^N p(B_n, D_n|A_n, D_n, \epsilon_n, v_n, \theta^{old})$

F. $\prod_n^N p(B_n, D_n|A_n, D_n, \theta^{old})$

G. None of the other answers is correct.

Solution: $\prod_n^N p(A_n|B_n, D_n, \theta^{old})$. Since B_n, D_n are observed, ϵ_n, v_n, and C_n are safely marginalized out, the only unobserved (latent) variable is A_n. Then, the posterior should be $p(A_{1:N}|B_{1:N}, D_{1:N}, \theta^{old}) = \prod_n^N p(A_n|B_n, D_n, \theta^{old})$.

Problem 33. [3 points] Given observations $\mathcal{O} = \{B_n, D_n\}_{n=1}^N$ and parameters θ. What is the variance of the conditional $p(A_n|\mathcal{O})$?

A. $[1 + \sigma_v^{-2} r^2 + \sigma_\epsilon^{-2} \sum_{k=1}^K w^2]^{-1}$

B. $1 + \sigma_v^{-2} r^2 + \sigma_\epsilon^{-2} \sum_{k=1}^K w^2$

C. $[\sigma_v^{-2} r^2 + \sigma_\epsilon^{-2} \sum_{k=1}^K w^2]^{-1}$

D. $1 + \sigma_\epsilon^{-2} r^2 + \sigma_v^{-2} \sum_{k=1}^K w^2$

E. $[\sigma_\epsilon^{-2} + \sigma_v^{-2} \sum_{k=1}^K w^2]^{-1}$

F. None of the other answers is correct.

Solution: $[1 + \sigma_v^{-2} r^2 + \sigma_\epsilon^{-2} \sum_{k=1}^K w^2]^{-1}$. Refer to Tutorial 6 question 1.c hint.

Problem 34. [2 points] Given observations $\mathcal{O} = \{B_n, D_n\}_{n=1}^N$, suppose we wish to learn θ (still deterministic parameters) by maximizing a variational lower-bound. As before, we first simplify the model by analytically marginalizing out ϵ_n, v_n, and C_n. The variational distribution q should be over which of the following sets of random variables? If multiple answers are correct, pick the one with the smallest number of random variables.

A. $\{A_n, B_n, C_n, D_n\}_{n=1}^N$

B. $\{A_n, B_n, C_n\}_{n=1}^N$

C. $\{A_n\}_{n=1}^N$

D. $\{A_n, D_n\}_{n=1}^N$

E. $\{C_n\}_{n=1}^N$

F. $\{B_n, C_n\}_{n=1}^N$

G. None of the other answers is correct.

Solution: C. $\{A_n\}_{n=1}^N$. Since B_n, D_n are observed, ϵ_n, v_n, and C_n are safely marginalized out, the only unobserved (latent) variable is A_n. Then, the variational distribution should be over $\{A_n\}_{n=1}^N$

Problem 35. [3 points] Given observations $\mathcal{O} = \{B_n, D_n\}_{n=1}^N$, suppose we wish to learn θ by optimizing a variational lower-bound. As before, we first simplify the model by analytically marginalizing out ϵ_n, v_n, and C_n and introduce a variational distribution q. Which of the following lower-bounds should we maximize given the model? The expectations are taken with respect to q

A. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n, C_n|rA_n + \mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big]$

B. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|rA_n + \mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big]$

C. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|rA_n + \mu, \sigma_v^2)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big]$

D. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|wA_n + \mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|rA_k, \sigma_\epsilon^2)]\Big]$

E. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|rA_n+\mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] - \mathbb{D}_{\text{KL}}[q(C_n)\|p(C_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big]$

F. None of the other answers is correct.

Solution: B. $\sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|rA_n + \mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big]$

$$\begin{aligned}
ELBO &= \sum_{n=1}^N \Big[\mathbb{E}_{q(A_n)}[\log \frac{p(A_n, B_n, D_n)}{q(A_n)}]\Big] && (12)\\
&= \sum_{n=1}^N \Big[\mathbb{E}_{q(A_n)}[\log p(D_n|A_n) \prod_{k=1}^K p(B_{n,k}|A_n) - \log q(A_n)]\Big] && (13)\\
&= \sum_{n=1}^N \Big[\mathbb{E}_{q(A_n)}[\log p(D_n|A_n)] + \sum_{k=1}^K \mathbb{E}_{q(A_n)}[\log p(B_{n,k}|A_n)] - \mathbb{E}_{q(A_n)}[\log \frac{q(A_n)}{p(A_n)}]\Big] && (14)\\
&= \sum_{n=1}^N \Big[\mathbb{E}_{q(A_n)}[\log p(D_n|A_n)] + \sum_{k=1}^K \mathbb{E}_{q(A_n)}[\log p(B_{n,k}|A_n)] - \mathbb{D}_{\text{KL}}(q(A_n)||p(A_n))\Big] && (15)\\
&= \sum_{n=1}^N \Big[\mathbb{E}[\log \mathcal{N}(D_n|rA_n + \mu, \sigma_v^2)] - \mathbb{D}_{\text{KL}}[q(A_n)\|p(A_n)] + \sum_{k=1}^K \mathbb{E}[\log \mathcal{N}(B_{n,k}|wA_k, \sigma_\epsilon^2)]\Big] && (16)
\end{aligned}$$

End of Paper