Homework #4

04654: Intro to Probabilistic Graphical Model
Prof. Assane GUEYE
Early deadline: Tuesday, Apr 29, 2025 at 11:59 PM CAT
Due: Thursday, May 01, 2025 at 11:59 PM CAT

Please remember to show your work for all problems and to write down the names of any students that you collaborate with. The full collaboration and grading policies are available on the course website: https://labayifa.github.io/04654. You are strongly encouraged (but not required) to use Latex to typeset your solutions.

Your solutions should be uploaded to Gradescope (https://www.gradescope.com/) in PDF format by the deadline. We will not accept hardcopies. If you choose to hand-write your solutions, please make sure the uploaded copies are legible. Gradescope will ask you to identify which page(s) contain your solutions to which problems, so make sure you leave enough time to finish this before the deadline. We will give you a 30-minute grace period to upload your solutions in case of technical problems.

1 Structure Learning *[50 points]*

1.1 Tree-Selection and the Chow-Liu Algorithm *[25 points]*

Use the Chow-Liu Algorithm to learn the model (tree and parameters) that generated the following data.

Data

1	0	0	1
0	1	0	0
1	1	0	1
1	1	1	1
0	1	0	1
1	0	0	1
1	1	0	1
0	0	1	0
0	1	0	1

1.2 Scoring Function *[25 points]*

Using the BIC scoring metrics below, compute the model (graph and potentials) that generated the data below. BIC:

$$S(G, \boldsymbol{\theta}; \mathbf{D}) = LL(\boldsymbol{\theta}; \mathbf{D}) - \varphi(|\mathbf{D}|)\|G\| \textbf{ where } \varphi(t) = \frac{\log(t)}{2}$$

Maximizing the score:

$$S_{max}(G, \mathbf{D}) = \max_{\boldsymbol{\theta}}(S(G, \boldsymbol{\theta}; \mathbf{D}))$$

Data

1	0	0
0	1	0
1	1	0
1	1	1
0	1	0
1	0	0
1	1	0
0	0	1
0	1	0

2 Variational Inference *[50 points]*

2.1 Mean-Field Approximation for Multivariate Gaussians *[40 points]*

In this question, we'll explore how accurate a Mean-Field approximation can be for an underlying multivariate Gaussian distribution. Assume we have observed data $\mathbf{X} \in \mathbb{R}^{2\times n}$ where each column $\mathbf{X}_{\cdot,i} \triangleq \mathbf{x}^{(i)} \in \mathbb{R}^2$ is a sample that was drawn from a 2-dimensional Gaussian distribution $\mathbf{x}^{(i)} \sim p(\cdot; \boldsymbol{\mu}, \ \boldsymbol{\Lambda}^{-1})$.

$$p(\mathbf{x}; \boldsymbol{\mu}, \ \boldsymbol{\Lambda}) = \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} \right) \tag{1}$$

Note here that we're using the *precision* matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. An additional property of the precision matrix is that it is symmetric, so $\Lambda_{12} = \Lambda_{21}$. (This is a convenient simplifying assumption.) We will approximate this 2-dimensional Gaussian with a mean field approximation, $q(\mathbf{x}) = q(x_1)q(x_2)$, the product of two 1-dimensional distributions $q(x_1)$ and $q(x_2)$. For now, we won't assume any form for this distributions. **Short Answer:** Write down the equation for $\log p(\mathbf{X})$. (For this question, you can leave all of the parameters in terms of vectors and matrices, not their subcomponents.)

1. (6 points)

2. (6 points) **Short Answer:** Group together everything that involves $\mathbf{X}_1$ and remove anything involving $\mathbf{X}_2$. We claim that there exists some distribution $q^*(\mathbf{X}) = q^*(\mathbf{X}_1)q^*(\mathbf{X}_2)$ that minimizes the KL divergence $q^* = \operatorname{argmin}_q \mathrm{KL}(q||p)$. And further, said distribution will have a component $q^\star(\mathbf{X}_1)$ will be proportional to the quantity you find below. Write that term that is proportional to $q^\star(\mathbf{X}_1)$.

It can be shown that this implies that $q(\mathbf{X}_1)$ (and therefore $q(\mathbf{X}_2)$) is a Gaussian distribution.

$$q(x_1) = \mathcal{N}(x_1; m_1, \Lambda_{11}^{-1})$$

Where $m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[x_2] - \mu_2)$

Using these facts, we'd like to explore how well our approximation can model the underlying distribution.

3. Suppose the parameters of the true distribution are $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 1/4 \end{pmatrix}$.

   (a) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(\mathbf{X}_1)$?

   (b) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(\mathbf{X}_1)$?

   (c) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(\mathbf{X}_2)$?

   (d) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(\mathbf{X}_2)$?

   (e) (5 points) **Plot:** Provide a *computer-generated* contour plot to show the result of our approximation $q^*(\mathbf{X})$ and the true underlying Gaussian $p(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ for the parameters given above.

4. Suppose the parameters of the true distribution are $\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}$.

(a) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(\mathbf{X}_1)$?

(b) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(\mathbf{X}_1)$?

(c) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(\mathbf{X}_2)$?

(d) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(\mathbf{X}_2)$?

(e) (5 points) **Plot:** Provide a *computer-generated* contour plot to show the result of our approximation $q^*(\mathbf{X})$ and the true underlying Gaussian $p(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ for the parameters given above.

5. (2 points) Describe in words how the plots you generated provide insight into the behavior of minimization of $KL(q||p)$ with regards to the low probability and high probability regions of the the true vs. approximate distributions.

2.2 Variational Inference vs. Monte Carlo Methods *[10 points]*

Let's end with a brief comparison between variational methods and MCMC methods. We have seen that both classes of methods can be used for learning in scenarios involving latent variables, but both have their own sets of advantages and disadvantages. For each of the following statements, specify whether they apply more suitably to VI or MCMC methods:

1. (2 points) Transforms inference into optimization problems.

   ◯ Variational Inference

   ◯ MCMC

2. (2 points) Is easier to integrate with back-propagation.

   ◯ Variational Inference

   ◯ MCMC

3. (2 points) Involves more stochasticity.

   ◯ Variational Inference

   ◯ MCMC

4. (2 points) Converges to the true distribution.

   ◯ Variational Inference

   ◯ MCMC

5. (2 points) Is higher variance under limited computational resources.

   ◯ Variational Inference

   ◯ MCMC