Course on probabilistic graphical models
Master MVA 2017-2018
Practice exercises 1: Solutions

These exercises are not meant to provide an exhaustive coverage of the material to review for the final exam. To some extend they focus more specifically on material that is not covered in the homeworks. Also, all these exercises should not be taken as representative of the difficulty of the questions posed at the exam, although several questions of the exam are likely to have a similar style. Some exercises are easy, and a few can be much harder so don't be discouraged if you find some of them difficult. They are primarily designed to help you review and consolidate your understanding of the course.

1. Let $p(x, y)$ be a joint probability distribution on two dependent random variables X and Y taking both values in $\{0, 1\}$

   (a) Write the joint distribution on (x, y) in exponential form

   $$\begin{aligned} p(x,y) &= \pi_{00}^{(1-x)(1-y)} \pi_{01}^{(1-x)y} \pi_{10}^{xy} \pi_{11}^{xy} \\ &= \exp\big((1-x)(1-y)\,\eta_{00} + (1-x)y\,\eta_{01} + x(1-y)\,\eta_{10} + xy\,\eta_{11}\big) \end{aligned}$$

   (b) What is the vector of sufficient statistics?

   $\phi(x,y) = \big((1-x)(1-y), (1-x)y, x(1-y), xy\big)^\top = \big((1-y), y\big)^\top \otimes \big((1-x), x\big)^\top$

   (c) Show that this distribution is in fact just a multinomial distribution

   With $z = (z_1, z_2, z_3, z_4) = \big((1-x)(1-y), (1-x)y, x(1-y), xy\big)$, *$z$ follows a usual multinomial distribution.*

   (d) Generalize to the joint distribution of n dependent Bernoulli random variables

   $\big((1-x_1), x_1\big) \otimes \big((1-x_2), x_2\big) \otimes \ldots \otimes \big((1-x_d), x_d\big)$ *follows a general multinomial distribution over* 2^d *different values.*

2. Consider a directed graphical model on the random variables $(X_1, \ldots, X_d)$ in which the likelihood takes the form

   $$p_\theta(x) = \prod_{j=1}^{d} p_{\theta_j}(x_j | x_{\pi_j}) \quad \text{with} \quad \theta = (\theta_1, \ldots, \theta_d) \in \Theta_1 \times \ldots \times \Theta_d.$$

   Assume that we have an i.i.d. sample of observations $x^{(1)}, \ldots, x^{(n)}$ with $x^{(i)} = \big(x_1^{(i)}, \ldots, x_d^{(i)}\big)$.

   (a) Explain why the maximum likelihood estimator can a priori be estimated easily. *Done in class. See lecture notes course 8.*

   (b) Why is it key to make the assumption that the set of possible values of the parameters is of the form $\Theta_1 \times \ldots \times \Theta_d$? *Because the optimization problem involving* $\theta_1, \ldots, \theta_d$ *decouple.*

   (c) Consider now the case in which the DAG is a tree and where X_j takes values in $\{0, 1\}$. Let $\theta_1 := \mathbb{P}(X_1 = 1)$ and, for $j > 1$, let $\theta_{j1} := \mathbb{P}(X_j = 1 | X_{\pi_j} = 1)$ and $\theta_{j0} := \mathbb{P}(X_j = 1 | X_{\pi_j} = 0)$;

let $\theta = (\theta_1, \theta_{21}, \theta_{20}, \ldots, \theta_{d1}, \theta_{d0})$ be parameterizing the model. What is the maximum likelihood estimator for θ based on the sample of size n introduced in the previous question?

The log-likelihood is $\ell(\theta) = \sum_{j=1}^{d} \ell_j(\theta_j)$ with

$$\begin{aligned}
\ell_j(\theta_j) &= \sum_{i=1}^{n} \log p(x_j^{(i)} \mid x_{\pi_j}^{(i)}) \\
&= \sum_{i=1}^{n} \big(x_j^{(i)} x_{\pi_j}^{(i)} \log \theta_{j1} + (1 - x_j^{(i)}) x_{\pi_j}^{(i)} \log(1 - \theta_{j1}) \ldots \\
&\quad \ldots + x_j^{(i)} (1 - x_{\pi_j}^{(i)}) \log \theta_{j0} + (1 - x_j^{(i)})(1 - x_{\pi_j}^{(i)}) \log(1 - \theta_{j0}) \big) \\
&= N_{j\pi_j,11} \log \theta_{j1} + N_{j\pi_j,01} \log(1 - \theta_{j1}) + N_{j\pi_j,10} \log \theta_{j0} + N_{j\pi_j,00} \log(1 - \theta_{j0})
\end{aligned}$$

with $N_{j\pi_j,kl}$ implicitly defined by the last equality. The maximum likelihood estimator is thus

$$\hat{\theta}_{j1} = \frac{N_{j\pi_j,11}}{N_{\pi_j}} \quad \text{and} \quad \hat{\theta}_{j0} = \frac{N_{j\pi_j,10}}{n - N_{\pi_j}},$$

with $N_j = \sum_{i=1}^{n} x_j^{(i)}$ for $j > 1$ and $\hat{\theta}_1 = \frac{N_1}{n}$.

(d) If we assume now that $\theta_{jk} = \theta_{j'k}$ for all $2 \leq j, j' \leq d$ and $k \in \{0, 1\}$, is it still possible to compute easily the maximum likelihood estimator? If so compute it.

The constraints introduce the coupling $\theta_2 = \theta_3 = \ldots = \theta_d$ so that we have:

$$\sum_{j=2}^{d} \ell_j(\theta_2) = \sum_{j=2}^{d} \big[N_{j\pi_j,11} \log \theta_{21} + N_{j\pi_j,01} \log(1 - \theta_{21}) + N_{j\pi_j,10} \log \theta_{20} + N_{j\pi_j,00} \log(1 - \theta_{20}) \big],$$

and thus for all $j \geq 2$,

$$\hat{\theta}_{j1} = \frac{\sum_{j=2}^{d} N_{j\pi_j,11}}{\sum_{j=2}^{d} N_{\pi_j}} \quad \text{and} \quad \hat{\theta}_{j0} = \frac{\sum_{j=2}^{d} N_{j\pi_j,10}}{n(d-1) - \sum_{j=2}^{d} N_{\pi_j}}.$$

3. Consider the joint Gaussian mixture distribution on $(X, Y) \in \mathbb{R}^2$ of the form:

$$p(x, y) = \frac{1}{4\pi} \Big(e^{-[x^2 + y^2]} + e^{-[(x-1)^2 + y^2]} + e^{-[(x-1)^2 + (y-1)^2]} + e^{-[x^2 + (y-1)^2]} \Big)$$

(a) Are X and Y correlated?

(b) Are X and Y independent?

(c) Propose a random variable Z such that $X \perp\!\!\!\perp Y | Z$

$p(x, y) = p_X(x) p_Y(y)$ with $p_X(x) = p_Y(x) = \frac{1}{2\sqrt{\pi}} \Big(e^{-x^2} + e^{-(x-1)^2} \Big)$ so clearly $X \perp\!\!\!\perp Y$ (and thus decorrelated) and any random variable Z independent of (X, Y) is such that $X \perp\!\!\!\perp Y | Z$.

Actually, there was a typo in the question, which was meant to consider the distribution:

$$p(x, y) = \frac{1}{4\pi} \Big(e^{-[x^2 + (y+1)^2]} + e^{-[x^2 + (y-1)^2]} + e^{-[(x-1)^2 + y^2]} + e^{-[(x+1)^2 + y^2]} \Big)$$

In that case it is natural to introduce the multinomial random variable $Z = (Z_1, Z_2, Z_3, Z_4)$ such that

$$\begin{aligned}
p(x, y | z_1 = 1) &= \frac{1}{\pi} e^{-[x^2 + (y+1)^2]} \\
p(x, y | z_2 = 1) &= \frac{1}{\pi} e^{-[x^2 + (y-1)^2]} \\
p(x, y | z_3 = 1) &= \frac{1}{\pi} e^{-[(x-1)^2 + y^2]} \\
p(x, y | z_4 = 1) &= \frac{1}{\pi} e^{-[(x+1)^2 + y^2]}.
\end{aligned}$$

Clearly $X \perp\!\!\!\perp Y \mid Z$ and we have $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ while we also have $\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\mathbb{E}[Y|Z] = 0$ because for any $k \in \{1, \ldots, 4\}$, either $\mathbb{E}[X|Z_k = 1] = 0$ or $\mathbb{E}[Y|Z_k = 1] = 0$. So that by linearity of expectations $\mathbb{E}[XY] = 0$. Altogether X and Y are decorrelated but they are not independent since

$$\mathbb{E}[X^2] = \frac{1}{4}\Big[\mathbb{E}[X^2|Z_1 = 1] + \ldots + \mathbb{E}[X^2|Z_4 = 1]\Big] = (1 + 1 + 2 + 2)/4 = 3/2 = \mathbb{E}[Y^2]$$

and $\mathbb{E}[X^2Y^2] = \mathbb{E}[X^2Y^2|Z] = 2 \neq \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$.

4. Consider X and Y two independent Bernoulli random variable. Let $Z = \max(X, Y) - XY$.

   (a) Which of the following statements are true: $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Y \mid Z$, $X \perp\!\!\!\perp Z \mid Y$, $Z \perp\!\!\!\perp Y \mid X$?

   The last three conditional independence statements clearly do not hold because Z is the exclusive or (XOR) of X and Y, so that any of the variables is exactly determined by the two others. For the pairwise independence statements: two cases need to be distinguished according to $p := \mathbb{P}(X = 1)$ and $q := \mathbb{P}(Y = 1)$. For general p and q none of the independence statement above hold. However for $p = q = \frac{1}{2}$ it can be immediately checked from probability tables that

$$\mathbb{P}(Z = 1|X = 1) = \mathbb{P}(Z = 1|X = 0) = \mathbb{P}(Z = 1|Y = 1) = \mathbb{P}(Z = 1|Y = 0) = \mathbb{P}(Z = 1) = \frac{1}{2},$$

   so that we have

$$X \perp\!\!\!\perp Y, \quad X \perp\!\!\!\perp Z, \quad Y \perp\!\!\!\perp Z,$$

   i.e. the variables are pairwise independent. They are clearly not jointly independent.

   (b) What is the smallest undirected graphical model containing the joint distribution over (X, Y, Z)? Is this satisfactory?

   It is the complete graph, which is unsatisfactory for $p = q = \frac{1}{2}$ because the independence statement holding for the distribution are not represented by the graph.

5. Assume that $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$. Is it true that $X_1 \perp\!\!\!\perp X_2 \mid (Y_1, Y_2)$? Prove or disprove.

   We have $p(x_1, x_2, y_1, y_2) = p(x_1, y_1)\, p(x_2, y_2)$ so that

$$p(y_1, y_2) = \sum_{x_1, x_2} p(x_1, x_2, y_1, y_2) = \Big[\sum_{x_1} p(x_1, y_1)\Big]\Big[\sum_{x_2} p(x_2, y_2)\Big] = p(y_1)p(y_2).$$

$$p(x_1, x_2 \mid y_1, y_2) = \frac{p(x_1, x_2, y_1, y_2)}{p(y_1, y_2)} = \frac{p(x_1, y_1)\, p(x_2, y_2)}{p(y_1)p(y_2)} = p(x_1 \mid y_1)\, p(x_2 \mid y_2).$$

   Now we clearly have $p(x_1, y_1, y_2) = p(x_1, y_1)p(y_2)$ and since $p(y_1, y_2) = p(y_1)p(y_2)$ we have

$$p(x_1|y_1, y_2) = \frac{p(x_1, y_1)p(y_2)}{p(y_1)p(y_2)} = p(x_1|y_1).$$

   Symmetrically $p(x_2|y_1, y_2) = p(x_2|y_2)$. So

$$p(x_1, x_2 \mid y_1, y_2) = p(x_1 \mid y_1)\, p(x_2 \mid y_2) = p(x_1 \mid y_1, y_2)\, p(x_2 \mid y_1, y_2),$$

   and the statement is true.

6. Let $X_1, \ldots, X_d$ be independent discrete random variables. Let $H(X_1, \ldots, X_d)$ and $H(X_i)$ denote respectively the entropy of the joint distribution of $(X_1, \ldots, X_d)$ and the entropy of the marginal distribution of X_i. Show that

$$H(X_1, \ldots, X_d) = \sum_{j=1}^{d} H(X_i)$$

$$H(X_1,\dots,X_d) = -\mathbb{E}[\log p(X_1,\dots,X_n)] = -\mathbb{E}\Big[\log \prod_i p(X_i)\Big] = -\sum_i -\mathbb{E}\Big[\log p(X_i)\Big] = \sum_{j=1}^{d} H(X_i)$$

7. **EM *vs* gradient descent in latent variable models.**

   Consider a latent variable model, with two variables (X, H), where X is observed and H is an unobserved latent variable. Let $\mathcal{P}_\Theta = \{p_\theta(x,h),\ \theta \in \Theta\}$ be a model for the pair (X,H), and consider the problem of learning the parameters in the model from observations of X alone, with marginal likelihood $p_\theta(x) = \int p_\theta(x,h)\,dh$. We have seen in class the EM algorithm, whose principle is to iteratively maximize a complete expected log-likelihood of the form $\mathbb{E}_{q_t}[\log p_\theta(x,H)]$ for $q_t(h) = p_{\theta_t}(h|x)$. The motivation was that the marginal log-likelihood $\log p_\theta(x)$ is non-convex and therefore not so simple to optimize. But the fact that a function is non-convex does not make it impossible to use methods such as gradient descent as soon as the log-likelihood is differentiable. In this exercise, we consider that option.

   (a) Show that under reasonable assumptions, if $g(\theta) = \nabla_\theta \log p_\theta(x)$ then the gradient of the marginal log-likelihood at a current value θ_t, $g(\theta_t)$, is in fact equal to the expected value under some distribution q_t that you will specify of the gradient of the complete log-likelihood.

   For $q_t(h) = p_{\theta_t}(h|x)$, the functions $\theta \mapsto \log p_\theta(x)$ and $\theta \mapsto \mathbb{E}_{q_t}[\log p_\theta(x,H)] + H(q_t)$ are tangent, because the first function is above the other for all θ, they touch at the point $\theta = \theta_t$ and both function are differentiable.

   This entails that at $\theta = \theta_t$, we have

$$\begin{aligned} g(\theta) = \nabla_\theta \log p_\theta(x) &= \nabla_\theta\big(\mathbb{E}_{q_t}[\log p_\theta(x,H)] + H(q_t)\big) \\ &= \mathbb{E}_{q_t}[\nabla_\theta \log p_\theta(x,H)], \end{aligned}$$

   provided we can exchange differentiation and expectation. To be allowed to do this we need to apply the dominated converge theorem and a sufficient condition is to have a domination by a integrable function, which is the case in the interior of the domain for members of an exponential family.

   (b) Explain how you would use this to compute $g(\theta_t)$ and why the previous result suggests that a step of probabilistic inference can in general not be avoided in the algorithm to compute that gradient.

   Since the gradient of the full log-likelihood is expressed in terms of the same sufficient statistics as full log-likelihood itself, it strongly suggests that to compute the gradient it will be necessary to perform the same E-step as in the EM algorithm, to use gradient descent.

8. Consider a distribution p that factorizes according to a directed graph G. Write p as an explicit product of potentials that show that p factorizes also with respected to the moralized graph associated to G.

   Take the conditionals as factors.

9. Let $p(x_1,\dots,x_d)$ be a distribution that factorizes according to an undirected tree with node set V and edge set E. Let $p(x_i,x_j)$ denote the marginal distribution over the pair of variables (X_i,X_j) and $p(x_i)$ be the marginal distribution over the variable X_i.

   (a) Show that

$$p(x_1,\dots,x_d) = \prod_{i\in V} p(x_i) \prod_{\{i,j\}\in E} \frac{p(x_i,x_j)}{p(x_i)p(x_j)}.$$

   We prove the result by induction. First, the statement is clearly true for a single node and a tree with a single edge. Assume that the nodes are ordered in a topological order. Make the induction

hypothesis that the result is true for a tree with $d-1$ nodes. Clearly, since d is a leaf, for the distribution considered, we have

$$p(x_1, \ldots, x_d) = p(x_1, \ldots, x_{d-1}) p(x_d | x_{\pi_d}).$$

and the marginal distribution $p(x_1, \ldots, x_{d-1})$ also factorizes according to a tree. Thus

$$p(x_1, \ldots, x_d) \quad = \quad \left[\prod_{i \in V \setminus \{d\}} p(x_i) \prod_{\{i,j\} \in E \setminus \{d, \pi_d\}} \frac{p(x_i, x_j)}{p(x_i) p(x_j)} \right] \cdot \frac{p(x_d, x_{\pi_d})}{p(x_{\pi_d}) p(x_d)} \cdot p(x_d).$$

By induction the result is proved.

(b) What is the exponential family form of the distribution if all variables X_i take values in $\{0, 1\}$?
After a simplification into a minimal representation in the exponential family, we obtain

$$p(x_1, \ldots, x_d) \quad = \quad \exp \Big[\sum_{i \in V} \eta_i x_i + \sum_{\{i,j\} \in E} \eta_{ij} x_i x_j - A(\eta) \Big],$$

because x_i, $(1 - x_i)$, $x_i x_j$, $(1 - x_i) x_j$, $x_i (1 - x_j)$, $(1 - x_i)(1 - x_j)$ are all linear combinations of 1, x_i, x_j and $x_i x_j$.