# Course on probabilistic graphical models
# Master MVA 2015-2016
# Brief solution of the first part of the exercises

December 15, 2015

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 was not covered in the homeworks and that corresponds to the four last lectures. Also, 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. So don't be discouraged if you find some of them difficult, in particular the problem, entitled trees, entropies and polytope, which is slightly harder. They are primarily designed to help you review and consolidate your understanding of the course.

## Various exercises

1. Consider a real-valued Gaussian homogeneous Markov chain specified by the recurrence $X_{t+1} = \rho X_t + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assuming that it is a Gaussian distribution, find the stationary distribution of this Markov chain, that is the distribution $P$ such that if $\mathbb{P}(X_1 \in A) = P(A)$ then $\mathbb{P}(X_t \in A) = P(A)$.

   *Surprisingly, nobody asked whether there was something missing in the text of the exercise. Two things are missing: $\epsilon$ is assumed to be independent of $X_t$ and $\epsilon$ should in fact be $\epsilon_t$ and these assumed i.i.d. Without these assumption and modification, something is missing in the problem and the problem is really weird. So let's make these changes.*

   *We want the distribution of $X_1$ to be the same as that of $X_0$ via a choice of the distribution of $X_0$. If the marginal distribution on $X_0$ is a stationary one and if we assume that it is Gaussian, it is characterized by its mean and variance.*

   *Let compute them: we must have $\mu = \mathbb{E}[X_1] = \rho\mathbb{E}[x_0] = \mu$ so that either $\rho = 1$ and there is no constraint on $\mu$ or $\rho \neq 1$ and then $\mu = 0$.*

   *For the variance, since $\epsilon_t \perp\!\!\!\perp X_t$, then $\tau^2 = Var(X_1) = \rho^2\, Var(X_0) + \sigma^2 = \rho^2\tau^2 + \sigma^2$. So if $\rho^2 > 1$ or $(\rho^2 = 1 \,\&\, \sigma^2 \neq 0)$, then there is no solution, i.e. no stationary distribution. If $\rho < 1$ then $\tau^2 = \sigma^2/(1-\rho^2)$ is a solution.*

   *To summarize: if $\rho^2 > 1$ or $(\rho^2 = 1 \,\&\, \sigma^2 \neq 0)$, then there is no stationary distribution. If $\rho^2 < 1$, then the stationary distribution is unique and it is $\mathcal{N}(0, \sigma^2/(1-\rho^2))$.*

2. 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

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

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

3. **Online Bayesian computations.** Let $(x_1,\dots,x_n)$ be a sequence of data observations. Let $\mathcal{P}_\Theta = \{p(x|\theta),\, \theta \in \Theta\}$ be a model for each of these observations and $\pi_0(\theta)$ a prior density on $\Theta \subset \mathbb{R}^d$. Define recursively

   $$\pi_t(\theta) = \frac{p(x_t|\theta)\,\pi_{t-1}(\theta)}{p_t(x_t)} \qquad \text{with} \qquad p_t(x_t) := \int p(x_t|\theta)\,\pi_{t-1}(\theta)\,d\theta.$$

   Prove that $\pi_n(\theta)$ is the posterior distribution given all the data $\pi_n(\theta) = p(\theta|x_1,\dots,x_n)$.

   *Proof by induction: for any $1 \le k \le n$, $\pi_k(\theta) = p(\theta|x_1,\dots,x_k)$*

4. **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 $\theta_t$, $g(\theta_t)$, 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 $\theta$, 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.*

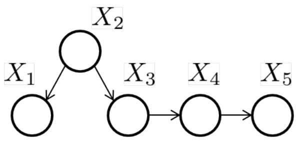
   (b) Explain how you would use this to compute $g(\theta_t)$ and why the previous result suggests that a step of probabilistic inference can in general not be avoided in the algorithm to compute that gradient.

   *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.*

5. 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.*

6. Consider running the sum-product algorithm on the following directed graphical model (DGM), where we suppose that each random variable takes value in $\{1,\dots,K\}$:



Express all answers as functions of the conditional probabilities $p(x_i|x_{\pi_i})$ for all $i$.

   (a) What is the message $m_{1\to 2}(x_2)$ sent from node $X_1$ to node $X_2$ during sum-product? Gives its simplest form.

   *Equal to 1 because corresponding to the marginalization of a leaf.*

   (b) Suppose that we observe the value of $X_3=\bar{x}_3$ and that we want to compute $p(x_2|\bar{x}_3)$. Gives the message $m_{3\to 2}(x_2)$ sent from node $X_3$ to node $X_2$.

$$\mu_{3\to 2}(x_2)=\sum_{x_3} p_{3|2}(x_3|x_2)\delta(x_3,\bar{x}_3)\mu_{4\to 3}(x_3),$$

   *but $\mu_{4\to 3}(x_3)=1$ for all values of $x_3$ for the same reason as in the previous question and so $\mu_{3\to 2}(x_2)=p_{3|2}(\bar{x}_3|x_2)$. Note however that even if $\mu_{4\to 3}(x_3)$ was a full message it would not have mattered here because of the conditioning on $x_3=\bar{x}_3$ since $X_{1,2}\perp\!\!\!\perp X_{4,5}\mid X_3$. In fact, in that case the message $m_{3\to 2}$ would just be changed by a constant factor which would disappear at the final renormalization. One thing to remember is that, multiplying a whole message by a constant does not change the result of the algorithm, which can be useful for numerical implementations.*

   (c) Give the expression in terms of messages during sum-product for $p(x_2|\bar{x}_3)$, as well as its simplified form.

   *To compute the marginal $p(x_2|\bar{x}_3)$ the sum product algorithm computes $\tilde{p}(x_2)=\psi_2(x_2)\mu_{3\to 2}(x_2)\mu_{1\to 2}(x_2)$ and then renormalizes it so that $\tilde{p}$ sums to 1. Since here 2 is the root, we have $\psi_2(x_2)=p_2(x_2)$ so that $\tilde{p}(x_2)=p_2(x_2)p_{3|2}(\bar{x}_3|x_2)$ we thus have $\sum_{x_2}\tilde{p}(x_2)=p_3(\bar{x}_3)$ and so by renormalizing $\tilde{p}(x_2)$ we effectively apply Bayes rule to retrieve the correct conditional.*

# Trees, entropies and polytopes

In the entire problem, $[K]$ denotes the set $\{1,\dots,K\}$.

## Part A

Consider a distribution $p(x_1,\dots,x_d)$ that factorizes according to an undirected tree $T=(V,E)$. Without loss of generality, we will assume that $X_i$ is a binary indicator vector of dimension $K$, i.e. that $X_i$ takes values in $\{0,1\}^K\cap\{x\mid\sum_{k=1}^K x_k=1\}$.

1. Show that the joint distribution on $(X_1,\dots,X_d)$, i.e. $p(x_1,\dots,x_d)$, can be always be expressed as an explicit function of the marginal distributions $\big(p_i(x_i)\big)_{i\in V}$ and of the pairwise distributions $\big(p_{ij}(x_i,x_j)\big)_{\{i,j\}\in E}$ only. *See solution of question 2*

2. Give an expression of $p(x_1,\dots,x_d)$ as a function of $(p_i)_{i\in V}$ and $(p_{ij})_{\{i,j\}\in E}$ that does not rely on some arbitrary orientation of the tree. (You might have to guess what the general form is and prove that it is correct by induction).

   *One can show that for a tree we always have*

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

*The proof is by induction on the number of leaves, by stripping one leaf at a time.*

3. Deduce from the previous formula that the entropy $H(X_1,\dots,X_d)$ for a distribution that factorizes according to a tree can be expressed as a function of $H(X_i)$ for all $i\in V$ and of all $I(X_i,X_j)$ for all $\{i,j\}\in E$.

   *Taking the log and then expectation of the previous formula we get*

$$H(X_1,\dots,X_d)=\sum_{i\in V}H(X_i)-\sum_{\{i,j\}\in E}I(X_i,X_j).$$

4. Show that it can alternatively be expressed as a function of $H(X_i)$ for $i\in V$, $H(X_i,X_j)$ for $\{i,j\}\in E$ and of the degrees $d_i$ of all nodes.

   *Using that $I(X_i,X_j)=H(X_i)+H(X_j)-H(X_i,X_j)$ we get*

$$H(X_1,\dots,X_d)=\sum_{i\in V}(1-d_i)H(X_i)+\sum_{\{i,j\}\in E}H(X_i,X_j),$$

   *where $d_i$ is the degree of node $i$.*

5. If the distribution $p(x_1,\dots,x_d)>0$, show that it can be written in exponential family form. What are the elements of the vector $\phi(x)$ of sufficient statistics? What are the components of the vector $\eta$ of natural parameters?

$$\log p(x_1,\dots,x_n)=\sum_{iinV}\sum_{k\in[K]}\eta_{ik}x_{ik}+\sum_{\{i,j\}\in E}\sum_{k,l\in[K]}\eta_{ijkl}x_{ik}x_{jl}-A(\eta).$$

6. What are the components of the vector $\mu$ of moment parameters?

   *The vector of moment parameters is formed of the*

$$\mu_{ik}=\mathbb{P}(X_{ik}=1)\quad \textit{and the}\quad \mu_{ijkl}=\mathbb{P}(X_{ik}=1,X_{jl}=1),$$

   *respectively for all $i\in V,k\in[K]$ and for all $\{i,j\}\in E$ and $k,l\in[K]$.*

7. Show that given a vector of moment parameters $\mu$, the corresponding vector of natural parameters $\eta$ can be expressed in closed form as a simple function of $\mu$. Explicit the function. Given the answer to question 2, we can write the same distribution as

$$p(x_1,\dots,x_n)=\left(\prod_{i\in V}\prod_{k\in[K]}\mu_{ik}^{x_{ik}}\right)\left(\prod_{\{i,j\}\in E}\sum_{k,l\in[K]}\left(\frac{\mu_{ijkl}}{\mu_{ik}\mu_{jl}}\right)^{x_{ik}x_{jl}}\right).$$

   *So that we can identify*

$$\eta_{ik}=\log\mu_{ik}\qquad \textit{and}\qquad \eta_{ijkl}=\log\frac{\mu_{ijkl}}{\mu_{ik}\mu_{jl}}.$$

8. Show that the entropy $H(X_1,\dots,X_d)$ can be expressed as a function of $\mu$ only. Explicit that function.

$$\widetilde{H}(\mu)=-\sum_{i\in V}\sum_{k\in[K]}\mu_{ik}\log\mu_{ik}-\sum_{\{i,j\}\in E}\sum_{k,l\in[K]}\mu_{ijkl}\log\frac{\mu_{ijkl}}{\mu_{ik}\mu_{jl}}.$$

## Part B

We consider the graphical model $\mathcal{P}_G$ consisting of all distributions over $\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ with $|\mathcal{X}_i| = K_i < \infty$ that factorize according to some undirected graph $G = (V, E)$. For convenience, we assume that $K_i = K$ for all $i$ and identify $\mathcal{X}_i$ with a copy of $\{0,1\}^K \cap \big\{x \mid \sum_{k=1}^K x_k = 1\big\}$.

Consider the set $\mathcal{H}_G$ defined by

$$\mathcal{H}_G := \big\{\mu = \big((\mu_i)_{i \in V}, (\mu_{ij})_{\{i,j\} \in E}\big)^\top \mid \mu_i \in [0,1]^K, \mu_{ij} \in [0,1]^{K \times K}\big\}.$$

The *local polytope* of $\mathcal{P}_G$ is the polytope $\mathcal{L}_G$ defined by

$$\mathcal{L}_G := \Big\{\mu \in \mathcal{H}_G \mid \forall \{i,j\} \in E, \sum_{l=1}^K [\mu_{ij}]_{kl} = [\mu_i]_k, \sum_{k=1}^K [\mu_{ij}]_{kl} = [\mu_j]_l\Big\}.$$

The *moment polytope* (or *marginal polytope*) is the set $\mathcal{M}_G$ defined by

$$\mathcal{M}_G = \big\{\mu \in \mathbb{R}^{Kd + K^2(d-1)} \mid \exists p \in \triangle_{\mathcal{X}}, \quad \mathbb{E}_p[\phi(X)] = \mu\big\},$$

where $\triangle_{\mathcal{X}}$ is the set of all probability distributions over $\mathcal{X}$, and

$$\phi(x) = \big((x_{ik})_{i \in V, k \in [K]}, (x_{ik}x_{jl})_{\{i,j\} \in E,\, k,l \in [K]}\big)^\top.$$

1. Show that for any graph $G$, we have $\mathcal{M}_G \subset \mathcal{L}_G$.

   *Any $\mu \in \mathcal{M}_G$ must satisfy the marginalization constraints, which shows the result.*

2. If $G = T$ is a tree, then use the results of part A to show that for any element $\tau \in \mathcal{L}_T$ such that all components of $\tau$ are strictly positive, it is possible to construct a distribution in $\mathcal{P}_T$ whose moments are equal to $\tau$.

   *Set $\mu = \tau$ in the first expression of the solution of question 7. This shows that there exists a tree distribution with exactly $\tau$ as a vector of moments.*

3. Use the result of the previous question to show that for a tree $T$, we must have $\mathcal{M}_T = \mathcal{L}_T$.

   *The moments that are strictly positive are dense in the set of all moments, and so the inclusion*

   $$\mathcal{L}_T \subset \mathcal{M}_T$$

   *follows from the previous question by closure.*