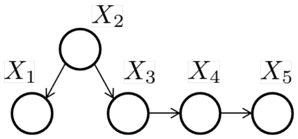
Course on probabilistic graphical models
Master MVA
Practice exercises 2

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.

Sum-product in a directed tree

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



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

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

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

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

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

   but $m_{4\to 3}(x_3) = 1$ for all values of x_3 for the same reason as in the previous question and so $m_{3\to 2}(x_2) = p_{3|2}(\bar{x}_3|x_2)$. Note however that even if $m_{4\to 3}(x_3)$ 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.

3. 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 $\tilde{p}(x_2) = \psi_2(x_2) m_{3\to 2}(x_2) m_{1\to 2}(x_2)$ 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.

Short problems

1. **Stationary distribution.** Consider a real-valued Gaussian homogeneous Markov chain specified by the recurrence $X_{t+1} = \rho X_t + \epsilon_t$ with $\epsilon_t \perp\!\!\!\perp X_t$ and $(\epsilon_t)_t$ an i.i.d. sequence following the normal distribution $\mathcal{N}(0, \sigma^2)$. 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 $\mathbb{P}(X_t \in A) = P(A)$.

   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's 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 μ 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. **Max entropy.** What is the family of distribution $(p_\alpha)_{\alpha \in \mathbb{R}_+}$ with p_α the distribution of maximal entropy on $\mathbb{N}$, the set of natural integers $\{0, 1, 2, \ldots\}$, such that $\mathbb{E}_p[X] = \alpha$?

   We have proved that the maximum entropy distribution with moment constraints on $\phi(x)$ is the distribution in the exponential family with sufficient statistic ϕ which maximizes the likelihood. In particular the constraint on the expectation of X means that $p(x) = \frac{1}{Z}e^{\eta x} = \frac{1}{Z}e^{\eta x} = \frac{1}{Z}\rho^x$ for $\rho = e^\eta$. We recognize the family of geometric distributions, with

   $$Z = \sum_{k=0}^{\infty} \rho^k = (1-\rho)^{-1}.$$

   We have

   $$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\rho^k = \rho \frac{\partial}{\partial \rho}\Big(\sum_{k=0}^{\infty} \rho^k\Big) = \frac{\rho}{1-\rho}.$$

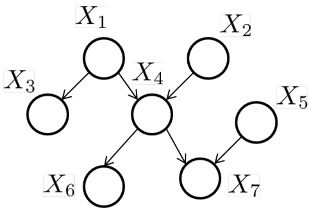
   which provides the relationship $\alpha = \frac{\rho}{1-\rho}$ or $\rho = \frac{\alpha}{1+\alpha}$.

3. **Sampling.** Propose a sampling scheme to sample exactly from the distribution $\mathbb{P}(X \in \cdot \mid \|X - y\|_2 \leq 1)$ where $y \in \mathbb{R}^d$ and X is a multivariate Gaussian random variable $\mathcal{N}(0, I_d)$. Prove that the proposed sampling scheme indeed yields a variable that has exactly the desired distribution.

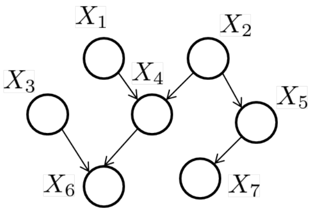
   We can use a rejection sampling scheme: sample from the Gaussian distribution and reject if the constraint is violated.

   Let's prove that it is indeed a rejection sampling scheme: let p_0 denote the Gaussian density, and p denote the density which we want to sample from. We have $p(x) = \frac{1}{Z}p_0(x)1_{\|X-y\|_2 \leq 1}$ as a consequence $p \leq \frac{1}{Z}p_0$, we can use $k = \frac{1}{Z}$ and $q = p_0$ for the proposal distribution and we have $p(x) \leq kq(x)$. The acceptance probability is $\frac{p(x)}{kq(x)}$ which is exactly 1 if the constraint is satisfied and zero else.

4. **Factorization.** Consider the two directed graphical models below. Is it possible to have a distribution p on $X_1, \ldots, X_7$ such that $p \in \mathcal{L}(G_1)$ and $p \in \mathcal{L}(G_2)$? Justify your answer.



G_1



G_2

Yes: the fully independent distribution ! The distributions in $\mathcal{L}(G)$ have to factorize according to the graph but they can certainly factorize more. The independent distribution is in all $\mathcal{L}(G)$ for all graphs G over a given number of nodes.

To give

Message passing for a Gaussian graphical model

Consider the directed Gaussian graphical models with the graph $G = (V, E)$ with $V = \{1, 2, 3, 4\}$ and $E = \big\{(1, 2), (1, 3), (2, 4)\big\}$ with the conditional probability densities $p(x_1) \propto \exp\big(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\big)$ and for all $(i, j) \in E$, $p(x_j|x_i) \propto \exp\big(-\frac{1}{2\sigma_j^2}(x_j - \rho_j x_i - \mu_j)^2\big)$.

We consider the problem of computing the distributions $p(x_1|x_3, x_4)$ and $p(x_2|x_3, x_4)$ using the sum-product algorithm in the graph in which the variables x_3 and x_4 are fixed to some given values.

1. Show that the sum-product algorithm consists in exchanging messages $\mu_{i\to j}(x_j)$ that are functions of the variable x_j and computed as integrals.

   The sum-product algorithm extends naturally to the case where we have densities. In that case, the messages are computed recursively with the formula

   $$m_{j\to i}(x_i) = \int_{x_j} \psi_{ij}(x_i, x_j)\psi_j(x_j)\Big(\prod_{k\in\mathcal{N}_j\setminus\{i\}} m_{k\to j}(x_j)\Big)dx_j.$$

   Clearly the message $m_{j\to i}$ is a function of x_i.

2. Show that the messages are of the form $m_{j\to i}(x_i) \propto \exp\frac{1}{2}\Big[\lambda_{j\to i}\, x_i^2 - 2\,\eta_{j\to i}\, x_i\Big]$.

   Since the potentials are Gaussian we could write generally

   $$\psi_i(x_i) = c_i \exp -\frac{1}{2}\Big[\lambda_{ii}^{(i)} x_i^2 - 2\eta_i^{(i)} x_i\Big]$$

   and

   $$\psi_{ij}(x_i, x_j) = c_{ij} \exp -\frac{1}{2}\Big[\lambda_{jj}^{(ij)} x_j^2 + \lambda_{ii}^{(ij)} x_i^2 + 2\lambda_{ij}^{(ij)} x_i x_j - 2\eta_i^{(ij)} x_i - 2\eta_j^{(ij)} x_j\Big].$$

   but this would be unnecessarily complicated: to avoid redundancy between the unary and the binary terms we will use a decomposition of the form

   $$\psi_i(x_i) = c_i \exp -\frac{1}{2}\Big[\lambda_{ii} x_i^2 - \eta_i x_i\Big] \qquad \text{and} \qquad \psi_{ij}(x_i, x_j) = c_{ij} \exp -\frac{1}{2}\Big[2\lambda_{ij} x_i x_j\Big],$$

   where only the cross-terms $x_i x_j$ are in the binary potentials. Clearly for any Gaussian distribution it is always possible to define the potentials this way.

   We show by induction that the messages are necessarily of the form

   $$m_{j\to i}(x_i) \propto \exp -\frac{1}{2}\Big[\lambda_{j\to i}\, x_i^2 - 2\,\eta_{j\to i}\, x_i\Big].$$

Note that it is perfectly fine to transmit messages up to a multiplicative constant that we ignore and do not try to compute, since it only changes the other messages by a multiplicative constant, and once marginals are computed up to a constant, finding the corresponding constant is easy by enforcing that the marginal distribution integrate to 1.

If we assume that the messages received by node j from all its neighbors different from i are of that form then, we have, using the formula

$$m_{j\to i}(x_i) = \int_{-\infty}^{+\infty} \psi_{ij}(x_i,x_j)\,\psi_j(x_j)\Big(\prod_{k\in\mathcal{N}_j\backslash\{i\}} m_{k\to j}(x_j)\Big)dx_j,$$

and the form of the Gaussian partition function $\int_{-\infty}^{+\infty}\exp\big(-\frac{1}{2}(\lambda y^2-2\eta y)\big)\,dy = \exp\big(\frac{1}{2}\frac{\eta^2}{\lambda}\big)$, that

$$m_{j\to i}(x_i) \propto \int_{-\infty}^{+\infty}\exp\bigg(-\frac{1}{2}\Big[\bar\lambda_j x_j^2 + 2\lambda_{ij}x_ix_j - 2\bar\eta_j x_j\Big]\bigg)dx_j \propto \exp\bigg(+\frac{1}{2}\frac{(\bar\eta_j-\lambda_{ij}x_i)^2}{\bar\lambda_j}\bigg),$$

with appropriate definitions of $\bar\lambda_j$ and $\bar\eta_j$, that is,

$$\bar\lambda_j = \lambda_j + \sum_{k\in\mathcal{N}(i)\backslash\{j\}}\lambda_{k\to j} \qquad \text{and} \qquad \bar\eta_j = \eta_j + \sum_{k\in\mathcal{N}(i)\backslash\{j\}}\eta_{k\to j}.$$

3. Show that instead of passing messages that are functions, it is sufficient to exchange messages that consist of the corresponding mean and variances.

   The formulas established in the previous question show we can just propagate the scalar messages $\lambda_{j\to i} = \frac{-\lambda_{ij}^2}{\bar\lambda_j}$ and $\eta_{j\to i} = -\frac{\bar\eta_j\lambda_{ij}}{\bar\lambda_j}$.

   (Note that $\lambda_{j\to i}$ alone cannot be interpreted as a variance (in particular it is negative), but that $\lambda_{ii}+\lambda_{j\to i}\geq 0$ and is naturally interpreted as a variance. So while this recursion does not propagate means and variances it is equivalent to a slightly different message algorithm for which this is exactly true.).

4. Express the update formulas for these mean-and-variance messages in terms of the means and variances propagated at all nodes in all directions.

   With the parameterization provided we have

$$\lambda_{11} = \frac{1}{\sigma_1^2}+\frac{\rho_2^2}{\sigma_2^2}+\frac{\rho_3^2}{\sigma_3^2}, \quad \lambda_{22} = \frac{1}{\sigma_2^2}+\frac{\rho_4^2}{\sigma_4^2}, \quad \lambda_{12} = -\frac{\rho_2}{\sigma_2^2}, \quad \lambda_{13} = -\frac{\rho_3}{\sigma_3^2}, \quad \lambda_{24} = -\frac{\rho_4}{\sigma_4^2}$$

$$\eta_1 = \frac{\mu_1}{\sigma_1^2}-\frac{\rho_2\mu_2}{\sigma_2^2}-\frac{\rho_3\mu_3}{\sigma_3^2}, \quad \eta_2 = \frac{\mu_2}{\sigma_2^2}-\frac{\rho_4\mu_4}{\sigma_4^2},$$

   Based on the formulas proved in the previous section, we therefore have

$$m_{3\to 1}(x_1)\propto\psi_{13}(x_1,x_3) \quad \text{and} \quad m_{4\to 2}(x_2)\propto\psi_{24}(x_2,x_4),$$

   so that

$$\lambda_{3\to 1}=\lambda_{4\to 2}=0, \quad \eta_{3\to 1}=-\lambda_{13}x_3, \quad \eta_{4\to 2}=-\lambda_{24}x_4.$$

   Then applying the formulas for $\eta_{1\to 2}$ and $\eta_{2\to 1}$ and identifying the different terms we get:

$$\lambda_{1\to 2} = -\frac{\lambda_{12}^2}{\lambda_{11}+\lambda_{3\to 1}} = -\frac{\lambda_{12}^2}{\lambda_{11}}, \quad \eta_{1\to 2} = -\frac{\lambda_{12}}{\lambda_{11}+\lambda_{3\to 1}}(\eta_1+\eta_{3\to 1}) = -\frac{\lambda_{12}\eta_1-\lambda_{12}\lambda_{13}x_3}{\lambda_{11}},$$

and symmetrically

$$\lambda_{2\to1} = -\frac{\lambda_{12}^2}{\lambda_{22}+\lambda_{4\to2}} = -\frac{\lambda_{12}^2}{\lambda_{22}}, \quad \eta_{2\to1} = -\frac{\lambda_{12}}{\lambda_{22}+\lambda_{4\to2}}(\eta_2+\eta_{4\to2}) = -\frac{\lambda_{12}\eta_2 - \lambda_{12}\lambda_{24}x_4}{\lambda_{22}}.$$

5. Similarly express the computation needed to compute the mean and variance of $p(x_1|x_3,x_4)$ and $p(x_2|x_3,x_4)$ from the messages arriving at the nodes 1 and 2 respectively.

We have

$$p(x_1|x_3,x_4) \propto \int_{-\infty}^{+\infty} \psi_1(x_1)m_{3\to1}(x_1)\, m_{2\to1}(x_1) = \exp -\frac{1}{2}\big[(\lambda_{11}+\lambda_{2\to1})\, x_1^2 - 2[\eta_1+\eta_{3\to1}+\eta_{2\to1}]\, x_1\big],$$

(where we did not write the null messages) which yields, using $\mu = \lambda^{-1}\eta$,

$$\mathbb{E}[X_1|X_3=x_3, X_4=x_4] = \frac{\eta_1+\eta_{3\to1}+\eta_{2\to1}}{\lambda_{11}+\lambda_{2\to1}}$$

and, using $\sigma^2 = \lambda^{-1}$,

$$\mathrm{Var}[X_1|X_3=x_3, X_4=x_4] = (\lambda_{11}+\lambda_{2\to1})^{-1} = \big(\lambda_{11} - \frac{\lambda_{12}^2}{\lambda_{22}}\big)^{-1}$$

Symmetrically, we have

$$\mathbb{E}[X_2|X_3=x_3, X_4=x_4] = \frac{\eta_2+\eta_{1\to2}+\eta_{4\to2}}{\lambda_{22}+\lambda_{1\to2}}$$

and

$$\mathrm{Var}[X_2|X_3=x_3, X_4=x_4] = (\lambda_{22}+\lambda_{1\to2})^{-1} = \big(\lambda_{22} - \frac{\lambda_{12}^2}{\lambda_{11}}\big)^{-1}.$$