Final exam
Probabilistic graphical models
Master MVA 2015-2016

December 16th 2015

A - Short problems

1. 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 $\rho = \frac{\alpha}{1+\alpha}$.

2. 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.

B - Factorization and Markov properties

1. Find an undirected graphical model on four random variables X_1, X_2, X_3, X_4 that satisfies simultaneous the following conditions (a) and (b)

   (a) All distributions that factorize according to the model satisfy the following conditional independence statements:

   $$X_1 \perp\!\!\!\perp X_2 \mid (X_3, X_4), \quad X_3 \perp\!\!\!\perp X_4 \mid (X_1, X_2),$$

(b) There exist distributions that factorize according to the model and such that

$$X_1 \not\perp\!\!\!\perp X_3 \mid X_2, \quad X_1 \not\perp\!\!\!\perp X_3, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_1 \not\perp\!\!\!\perp X_4, \quad X_2 \not\perp\!\!\!\perp X_4.$$

There are 4 different solutions:

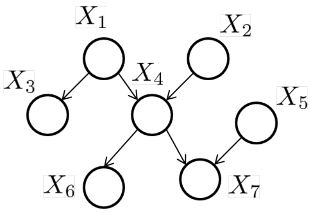
- *the cycle* $(1, 3, 2, 4)$*,*
- *the chain* $3 - 1 - 4 - 2$*,*
- *the chain* $1 - 3 - 2 - 4$*.*
- *the chain* $4 - 1 - 3 - 2$*.*

So see this note that the conditional independence statements in (a) impose that

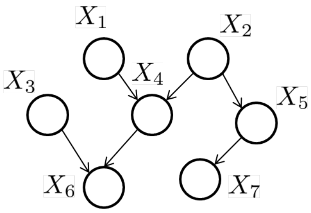
- *there is no edge* $\{1, 2\}$*,*
- *there is no edge* $\{3, 4\}$*.*

Then the condition that there exist a distribution such that $X_1 \not\perp\!\!\!\perp X_3 \mid X_2$ *implies that, when node 2 is removed, there is still a path connecting* 1 *and* 3*. But since there is no edge* $\{3, 4\}$ *there must be an edge* $\{1, 3\}$*. The remaining statements taken together impose that at least two out of the three remaining allowed edges (*$\{1, 4\}$*,*$\{2, 3\}$*,*$\{2, 4\}$*) need to be included. One can then check that all possible graphs obtained by adding two of these three edges or all three of them satisfy all the conditions.*

2. 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.*

C - EM algorithm

Consider (X, Y) a pair of correlated Bernoulli random variables which we parameterize by $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ with $\pi_{kj} = \mathbb{P}(X = k, Y = j)$. Assume that we observe the three following independent samples

- an i.i.d. sample $(X_1, \ldots, X_{n_x})$ from the marginal distribution of X, together with
- an i.i.d. sample $(Y_1, \ldots, Y_{n_y})$ from the marginal distribution of Y, and
- an i.i.d. sample $\big((X'_1, Y'_1), \ldots, (X'_m, Y'_m)\big)$ from the joint distribution of (X, Y)

Consider the notations:

- $N_x = \sum_{i=1}^{n_x} X_i,\ N_y = \sum_{i=1}^{n_y} Y_i,$
- $M_{jk} = \sum_{i=1}^{m} \delta(X'_i, k)\, \delta(Y'_i, j)$

Propose an EM algorithm to estimate $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$. In particular, specify which quantities are computed for the E-step and which quantities are computed for the M-step.

We denote $\tilde{y}_i$ the variable Y associated with x_i which is not observed and $\tilde{x}_i$ the variable X associated with y_i which is not observed

The full log-likelihood is composed of three terms

$$A = \sum_{i=1}^{n_x} \left[x_i\, \tilde{y}_i \log \pi_{11} + \overline{x_i}\, \tilde{y}_i \log \pi_{01} + x_i\, \overline{\tilde{y}_i} \log \pi_{10} + \overline{x_i}\, \overline{\tilde{y}_i} \log \pi_{00} \right]$$

$$B = \sum_{i=1}^{n_y} \left[\tilde{x}_i\, y_i \log \pi_{11} + \overline{\tilde{x}_i}\, y_i \log \pi_{01} + \tilde{x}_i\, \overline{y_i} \log \pi_{10} + \overline{\tilde{x}_i}\, \overline{y_i} \log \pi_{00} \right]$$

$$C = \sum_{i=1}^{m} \left[x_i\, y_i \log \pi_{11} + \overline{x_i}\, y_i \log \pi_{01} + x_i\, \overline{y_i} \log \pi_{10} + \overline{x_i}\, \overline{y_i} \log \pi_{00} \right]$$

which can also we written

$$C = M_{11} \log \pi_{11} + M_{01} \log \pi_{01} + M_{10} \log \pi_{10} + M_{00} \log \pi_{00}.$$

Let q be the conditional distribution of the unobserved data given the observed data at time t. For now we don't write the dependence on t. We denote

$$q_{y|x=0} = \mathbb{P}_\theta(Y = 1|X = 0), \quad q_{y|x=1} = \mathbb{P}_\theta(Y = 1|X = i),$$
$$q_{x|y=0} = \mathbb{P}_\theta(X = 1|Y = 0), \quad q_{x|y=1} =_\theta \mathbb{P}(X = 1|Y = i).$$

We have

$$\mathbb{E}_q[A] = N_x q_{y|x=1} \log \pi_{11} + N_x(1 - q_{y|x=1}) \log \pi_{10} + (n_x - N_x) q_{y|x=0} \log \pi_{01} + (n_x - N_x)(1 - q_{y|x=0}) \log \pi_{00}.$$

$$\mathbb{E}_q[B] = N_y q_{x|y=1} \log \pi_{11} + N_y(1 - q_{x|y=1}) \log \pi_{01} + (n_y - N_y) q_{x|y=0} \log \pi_{10} + (n_y - N_y)(1 - q_{x|y=0}) \log \pi_{00}.$$

With these calculations we can specify the EM algorithm: At time t, given the current parameters values $(\pi^t_{00}, \pi^t_{01}, \pi^t_{10}, \pi^t_{11})$ we have

$$q^t_{y|x=0} = \mathbb{P}_{\theta^t}(Y = 1|X = 0) = \frac{\pi^t_{01}}{\pi^t_{00} + \pi^t_{01}}, \quad q^t_{y|x=1} = \mathbb{P}_{\theta^t}(Y = 1|X = 1) = \frac{\pi^t_{11}}{\pi^t_{10} + \pi^t_{11}}$$

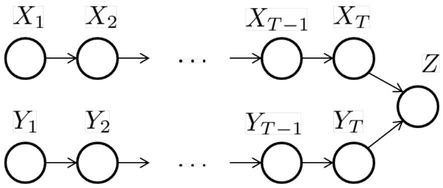
$$q^t_{x|y=0} = \mathbb{P}_{\theta^t}(X = 1|Y = 0) = \frac{\pi^t_{10}}{\pi^t_{00} + \pi^t_{10}}, \quad q^t_{x|y=1} = \mathbb{P}_{\theta^t}(X = 1|Y = 1) = \frac{\pi^t_{11}}{\pi^t_{01} + \pi^t_{11}}$$

Now for the M-step we have

$$\begin{aligned}
\pi_{11} &= \frac{N_x\, q^t_{y|x=1} + N_y\, q^t_{x|y=1} + M_{11}}{n_x + n_y + m} \\
\pi_{10} &= \frac{N_x\, (1 - q^t_{y|x=1}) + (n_y - N_y)\, q^t_{x|y=0} + M_{10}}{n_x + n_y + m} \\
\pi_{01} &= \frac{(n_x - N_x)\, q^t_{y|x=0} + N_y\, (1 - q^t_{x|y=1}) + M_{01}}{n_x + n_y + m} \\
\pi_{00} &= \frac{(n_x - N_x)\, (1 - q^t_{y|x=0}) + (n_y - N_y)\, (1 - q^t_{x|y=0}) + M_{00}}{n_x + n_y + m}
\end{aligned}$$

D - Parallel chains

Consider the directed graphical model G below:



Suppose that the variables X_t and Y_t are discrete K-valued for $t = 1, \ldots, T$, and that Z is a binary random variable. Consider the following form for the factors for a specific distribution p in $\mathcal{L}(G)$:

- $p(X_1 = l) = p(Y_1 = l) = \pi_l$ for $l = 1, \ldots, K$.
- $p(X_t = i|X_{t-1} = j) = p(Y_t = i|Y_{t-1} = j) = A_{ij}$ for $i, j = 1, \ldots, K$ and $t = 2, \ldots, T$.
- $p(Z = 1|X_T = k, Y_T = l)$ takes the value p whenever $k = l$, and q otherwise (i.e. when $k \neq l$).

By using the graph eliminate algorithm (or just clever use of distributivity), give a simple formula for the marginal $p(Z = 1)$ in terms of the matrix A, vector π and scalars q, p and T. Provide brief explanations of how to obtain the result.

By recurrence the message received by Z by each of its neighbor is $m := A^{T-1}\pi$. The marginal probability that $Z = 1$ is

$$m^\top \big((p-q)\,\mathbf{I} + q\,\mathbf{1}\mathbf{1}^\top\big)\, m,$$

with $\mathbf{I}$ the identity matrix.

E - Metropolized Gibbs sampler

We consider the Gibbs distribution p of the random variable $X = (X_1, \ldots, X_n)$ (which we can think of as a distribution from a graphical model of interest). For simplicity we assume that the X_i are K-valued random variables. We will use the notation X_{-i} to refer to $X_{\{1,\ldots,n\}\backslash\{i\}}$. We will denote by p_i the conditional distribution of the ith variable given all the others, as induced from the Gibbs distribution so that

$$p_i(\, z_i \mid x^t_{-i}) := \mathbb{P}(X_i = z_i \mid X_{-i} = x^t_{-i}).$$

The Metropolized Gibbs sampler is a variant of Gibbs sampling, which takes the form of a Metropolis-Hasting algorithm that updates a single variable X_i at a time: to update the variable X_i, instead of just sampling this variable conditionally on the other variables, it makes a proposal drawn from the transition T_i with

$$T_i\big((x^t_i, x^t_{-i}), (z_i, x^t_{-i})\big) := \begin{cases} \dfrac{p_i(z_i \mid x^t_{-i})}{1 - p_i(x^t_i \mid x^t_{-i})} & \text{for } z_i \neq x^t_i \\ 0 & \text{for } z_i = x^t_i \end{cases} \tag{1}$$

and accepts the move with probability

$$\alpha_i((x^t_i, x^t_{-i}), (z_i, x^t_{-i})\big) := \min\left\{1, \frac{1 - p_i(x^t_i \mid x^t_{-i})}{1 - p_i(z_i \mid x^t_{-i})}\right\}.$$

You may use the notations $\alpha_i(x^t_i, z_i) := \alpha_i\big((x^t_i, x^t_{-i}), (z_i, x^t_{-i})\big)$ and $T_i(x^t_i, z_i) := T_i\big((x^t_i, x^t_{-i}), (z_i, x^t_{-i})\big)$ in you answers to simplify the notations.

1. Show that the transition corresponding to an update of variable X_i of the Metropolized Gibbs sampler satisfies detailed balance with the Gibbs distribution.

   We compute the transition probability of the M-G sampler for $z_i \neq x_i^t$:

   $$S_i(x_i^t, z_i) = T_i(x_i^t, z_i)\, \alpha_i(x_i^t, z_i) = p_i(z_i | x_{-i}^t) \, \min \Big(\frac{1}{1 - p_i(x_i^t \mid x_{-i}^t)}, \frac{1}{1 - p_i(z_i \mid x_{-i}^t)} \Big).$$

   Checking reversibility is only necessary for $z_i \neq x_i^t$ as it is obvious for $z_i = x_i^t$.

   $$\begin{aligned} p(x^t) S_i(x_i^t, x_i^{t+1}) &= p(x_{-i}^t)\, p_i(x_i^t | x_{-i}^t)\, S_i(x_i^t, x_i^{t+1}) \\ &= p(x_{-i}^t)\, p_i(x_i^t | x_{-i}^t)\, p_i(x_i^{t+1} | x_{-i}^t) \, \min \Big(\frac{1}{1 - p_i(x_i^t \mid x_{-i}^t)}, \frac{1}{1 - p_i(x_i^{t+1} \mid x_{-i}^t)} \Big) \\ &= p(x_{-i}^{t+1})\, p_i(x_i^t | x_{-i}^{t+1})\, p_i(x_i^{t+1} | x_{-i}^{t+1}) \, \min \Big(\frac{1}{1 - p_i(x_i^t \mid x_{-i}^{t+1})}, \frac{1}{1 - p_i(x_i^{t+1} \mid x_{-i}^{t+1})} \Big) \\ &= p(x_{-i}^{t+1})\, p_i(x_i^{t+1} | x_{-i}^{t+1})\, p_i(x_i^t | x_{-i}^{t+1}) \, \min \Big(\frac{1}{1 - p_i(x_i^t \mid x_{-i}^{t+1})}, \frac{1}{1 - p_i(x_i^{t+1} \mid x_{-i}^{t+1})} \Big) \\ &= p(x^{t+1}) S_i(x_i^{t+1}, x_i^t) \end{aligned}$$

   which proves the reversibility.

2. Does the Markov Chain produced by the Metropolized Gibbs sampler converge to the Gibbs distribution? Explain why. To answer this question you can either consider a random scan Gibbs sampler that picks the variable i uniformly at random at each iteration or a cyclic scan Gibbs sampler that samples each variable X_i in a cycle.

   The M-G transition for a single variable allow the transitions to all possible values of x_i if and only if $p_i(x_i | x_{-i}) > 0$ for all values of x_i and x_{-i}. Obviously, even if that condition is met, the M-G transition for a single variable does not allow to go from any configuration of $x \in \mathbb{R}^n$ to any other since a single scalar variable is modified at time. However, if we either make a whole cycle (in the case of the cyclic scan) or have at least n random sampling steps (in the case of the random scan) the probability of going from any configuration to any other is non zero.

   To reason about aperiodicity and irreducibility, let's distinguish between the cyclic scan and the random scan:

   - *In the case of a cyclic scan, the transition should be considered to be the whole cycle, and what we have just shown above is that that transition is regular which means that it is irreducible and aperiodic*
   - *In the case of a random scan, the transition corresponds to the update of a single coordinate. Let's denote by $S(x, y)$ this transition which consists in choosing a random coordinate and then updating it. We have argued above that $S^n(x, y) > 0$ for all (x, y). But since x_i can stay the same with non-zero probability in a single step, we have $S(x, x) > 0$ and so it is also true that $S^{n+k}(x, y) > 0$ for all (x, y) and for all $k \in \mathbb{N}$. This shows that this transition S is irreducible aperiodic.*

   To conclude, we have seen in class a main theorem that states that if a transition which is irreducible and aperiodic satisfies detailed balance for a certain distribution π, then the corresponding Markov chain converges in distribution to variable with distribution π provided the latter puts positive mass on all configurations. This proves the result.

F - Bayesian estimation: posterior mean *vs* MAP

Assume we observe an i.i.d. sample $(x_1, \dots, x_n)$ of realizations of a Bernoulli random variable whose unknown moment parameter we denote by $\mu = \mathbb{E}_\mu[X] = \mathbb{P}_\mu(X = 1)$.

We consider first the Bayesian estimation of μ and the corresponding MAP estimator, based on a uniform prior distribution of μ.

1. Express the posterior mean estimate $\hat{\mu}_{\mathrm{PM}}$ for μ as a function of $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\hat{\mu}_{PM} = \frac{\bar{x}_i + 1}{n+2}$$

2. What is the maximum a posteriori (MAP) estimate $\hat{\mu}_{\mathrm{MAP}}$ for μ?

   The same as the maximum likelihood estimator: $\bar{x}$.

We now consider again the same estimators, but we instead work with η, the canonical parameter of the exponential family.

(c) If a uniform prior distribution $p_\mu(\mu)$ is placed on μ what is the induced prior distribution on η? Provide the density of that prior $p_\eta(\eta)$ when p_μ is the uniform prior on the interval $[0,1]$. *We have*

$$\eta = \log \frac{\mu}{1-\mu}$$

or equivalently

$$\mu = \frac{e^\eta}{e^\eta + 1}.$$

Note that $\eta \mapsto \frac{e^\eta}{e^\eta+1}$ can be viewed as a cumulative density function, say F. But if μ is distributed uniformly on $[0,1]$ then η has the same distribution as $F^{-1}(U)$, where U is a uniform. This means that F is exactly the cdf of η. The corresponding pdf on $\mathbb{R}$ is

$$p_\eta(\eta) = \frac{e^\eta}{(1+e^\eta)^2} = \mu(\eta)\,\big(1-\mu(\eta)\big).$$

We recognize the logistic distribution, *which resembles the Gaussian distribution.*

(d) What is the value of the posterior mean estimate $\int \mu(\eta)\, p(\eta|x_1,\ldots,x_n)\, d\eta$ under this prior p_η on η, for $\mu(\eta) = (1+e^{-\eta})^{-1}$ the usual moment mapping?

The posterior mean does not change if we make a change of variable. We thus have:

$$\int \mu(\eta)\, p(\eta|x_1,\ldots,x_n)\, d\eta = \hat{\mu}_{PM} = \frac{\bar{x}_i + 1}{n+2}$$

(e) For this prior p_η, what is the MAP estimator $\hat{\eta}_{\mathrm{MAP}}$? What is the corresponding moment parameter $\mu(\hat{\eta}_{\mathrm{MAP}})$?

The calculations lead to the same result as for the PM estimator. Indeed the likelihood times the prior is

$$\Big(\prod_{i=1}^{n} p(x_i|\eta)\Big) p_\eta(\eta) = \mu(\eta)^{n\bar{x}}(1-\mu(\eta))^{n(1-\bar{x})} \cdot \mu(\eta)\,\big(1-\mu(\eta)\big)$$

(f) Comment on the different estimators obtained.

While the posterior mean does not depend on the parameterization, the MAP estimator does. This is one of the frailties of the MAP estimator.