1 Poisson distributions and multiplicative updates

We consider a random variable $y \in \mathbb{N}$ on non-negative integers, and the Poisson distribution with parameter $\lambda$, with density $p(y) = \frac{e^{-\lambda} \lambda^y}{y!}$ with respect to the counting measure.

1. (1 point) Show that this is a member of an exponential family and compute the log-partition function.

**Solution:** We have: $p(y) = \frac{1}{y!} \exp(y \log \lambda - \lambda) = \frac{1}{y!} \exp(y \eta - e^\eta)$ with $\eta = \log \lambda$ the natural parameter and $A(\eta) = e^\eta$ the log-partition function.

2. (2 points) Compute the mean and variance of a Poisson random variable using properties of exponential families.

**Solution:** We have $\mathbb{E} y = A'(\eta) = e^\eta = \lambda$ and $\text{var}(y) = A''(\eta) = e^\eta = \lambda$.

Poisson regression

We consider $y \in \mathbb{N}^n$ and $X \in \mathbb{R}_+^{n \times d}$, and the model $y_i \sim \text{Poisson}((X w)_i)$ for $w \in \mathbb{R}_+^d$ and all $y_i$, $i \in \{1, \ldots, n\}$ being independent.

3. (2 points) Write down the log-likelihood $\ell(w)$ of the model. Is it concave? Is it convex?
Solution: We have $\ell(w) = \sum_{i=1}^n \{y_i \log (Xw)_i - (Xw)_i - \log y_i!\}$. It is concave in $w$.

4. (2 points) By using Jensen’s inequality for the logarithm, show that $\ell(w)$ can be written in the form

$$\ell(w) = \sup_{Q \in \mathbb{R}^{n \times d}} \sum_{i=1}^n y_i \sum_{j=1}^d Q_{ij} \log \frac{X_{ij} w_j}{Q_{ij}} + \text{function}(w),$$

with the constraint that $\forall i \in \{1, \ldots, n\}$, $\sum_{j=1}^d Q_{ij} = 1$.

Solution: We have

$$\ell(w) = \sum_{i=1}^n \{y_i \log (Xw)_i - (Xw)_i - \log y_i!\}$$

$$= \sum_{i=1}^n y_i \log \sum_{j=1}^d Q_{ij} \frac{X_{ij} w_j}{Q_{ij}} - 1^\top Xw - \sum_{i=1}^n \log y_i!$$

$$\geq \sum_{i=1}^n y_i \sum_{j=1}^d Q_{ij} \log \frac{X_{ij} w_j}{Q_{ij}} - 1^\top Xw - \sum_{i=1}^n \log y_i! \text{ by Jensen’s inequality.}$$

5. (3 points) Write down the updates for an alternative maximization algorithm for $(w, Q)$. Where do they converge to?

Solution:
We have equality in Jensen’s inequality is $Q_{ij} = \frac{X_{ij} w_j}{(Xw)_j}$. Given $Q$, the part of the function that depends on $w_j$ is $\sum_{i=1}^n Q_{ij} y_i \log w_j - w_j (1^\top 1)_j$ leading to the update

$$w_i = \frac{\sum_{i=1}^n Q_{ij} y_i}{(1^\top 1)_j} = \frac{(Q^\top y)_j}{(1^\top 1)_j}.$$

If initialized to strictly positive components it converges to the global optimum.

Relationship with the EM algorithm

6. (2 points) Preliminary 1: show that the sum $y$ of two independent Poisson random variables $y_1$ and $y_2$, with parameters $\lambda_1$ and $\lambda_2$ is Poisson with parameters $\lambda_1 + \lambda_2$. Hint: compute the characteristic function $\varphi(\theta) = E e^{i\theta y}$ for $y = y_1 + y_2$.

Solution: The characteristic function of a Poisson random variable is $\sum_{y=0}^{\infty} p(y) e^{i\theta y} = \sum_{y=0}^{\infty} e^{-\lambda} \frac{\lambda^y}{y!} e^{i\theta y} = \exp(\lambda e^{i\theta} - \lambda)$. We have, by independence of $y_1$ and $y_2$, $\varphi(\theta) = E e^{i\theta y_1} E e^{i\theta y_2} = \exp(\lambda_1 e^{i\theta} - \lambda_1) \exp(\lambda_2 e^{i\theta} - \lambda_2) = \exp((\lambda_1 + \lambda_2) e^{i\theta} - (\lambda_1 + \lambda_2))$, which is the characteristic function of the desired random variable.

7. (2 points) Preliminary 2: If $y$ is the sum of two independent Poisson random variables $y_1$ and $y_2$, with parameters $\lambda_1$ and $\lambda_2$, show that $E(y_1 | y) = \frac{\lambda_1}{\lambda_1 + \lambda_2} y$. 

Page 2
Solution: We have:

\[
E(y_1 | y) = \frac{1}{p(y)} \sum_{x_1=0}^{y} p(y_1 = x_1 | y) x_1 = \frac{1}{p(y)} \sum_{x_1=1}^{y} p(y_1 = x_1, y_2 = y - x_1) x_1 = \frac{y!}{e^{-\lambda_1 - \lambda_2} (\lambda_1 + \lambda_2)^y} \sum_{x_1=1}^{y} \frac{x_1! e^{-\lambda_1 \lambda_1 x_1^y} e^{-\lambda_2 \lambda_2 y-x_1}}{(y-x_1)!} = \frac{y! \lambda_1 \lambda_1}{(\lambda_1 + \lambda_2)^y} \sum_{x_1=1}^{y} \frac{(\lambda_1/\lambda_2)^{x_1-1} (y-1)!}{(x_1-1)! (y-x_1)!} = \frac{y! \lambda_1 \lambda_1}{\lambda_1 + \lambda_2} \frac{(\lambda_1/\lambda_2 + 1)^{y-1}}{y}.
\]

8. (1 point) We consider the mutually independent \(nd\) latent variables \(z_{ij} \sim \text{Poisson}(X_{ij} w_j)\), for \(i \in \{1, \ldots, n\}\) and \(j \in \{1, \ldots, d\}\), and \(y_i = \sum_{j=1}^{d} z_{ij}\) for \(i \in \{1, \ldots, n\}\). What is the distribution of \(y_i\)?

Solution: This is Poisson with parameter \((X w)_i\).

9. (1 point) Write down the complete log-likelihood of the variables \(z_{ij}\), \(i \in \{1, \ldots, n\}\) and \(j \in \{1, \ldots, d\}\).

Solution: We have \(\mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{d} \{z_{ij} \log X_{ij} w_j - X_{ij} w_j - \log z_{ij}\}\).

10. (2 points) What are the expectations needed for the E-step of the EM algorithm? Compute them.

Solution: Using (b), we get \(\langle z_{ij} \rangle = E(z_{ij} | y_i) = y_i \frac{X_{ij} w_j}{(X w)_i}\).

11. (2 points) What is the M-step?

Solution: We need to maximize \(\sum_{i=1}^{n} \sum_{j=1}^{d} \{\langle z_{ij} \rangle \log X_{ij} w_j - X_{ij} w_j\}\), which leads to \(w_j = \frac{\sum_{i=1}^{n} \langle z_{ij} \rangle}{\sum_{i=1}^{n} X_{ij}}\).

12. (1 point) Compare the two algorithms from parts (2) and (3).

Solution: They are the same.
2 Crossroads

Consider a model represented by the directed graph above, where each node is a random variable taking values in the range 1, \ldots, K.

1. (1 point) What type of graph is this? (Explain)

\textbf{Solution:} This is a \textbf{directed tree}: one node (node 1) has no parent, all the other nodes have exactly one parent.

2. (2 points) What is the dimension of the parameter space? How would you learn the parameters of the model, given observations $X_1, \ldots, X_n$?

\textbf{Solution:} Let $\pi$, the distribution of the variable $X_1$ (root node); $\pi$ is a vector of $K$ probabilities, that sum to one, $\pi_1 + \ldots + \pi_K = 1$. The six other nodes has exactly one parent (again, this is a tree); so let $A_{ikl} = P(X_i = l | X_{pa(i)} = k)$; these probabilities also sum to one (w.r.t. $k$). In the end, $d = (K - 1)(1 + 6K)$.

To learn the parameters, simply compute the corresponding proportions; for instance, set $\hat{\pi}_k$ to the proportions of datasets where $X_1$ takes value $k$, and set $A_{ikl}$ to the proportion of datasets such that $X_i = l$ (among those datasets such that the parent of $X_i$ takes values $k$).

3. (1 point) Explain how you may sample observations from this model (given fixed parameters). Recall the name of the corresponding sampling algorithm.

\textbf{Solution:} We may use the \textbf{ancestral sampling} algorithm: first sample $X_1$ (according to the vector of probabilities $\pi$); call $x_1$ the result. Then we may sample $X_2$ according to the probabilities $A_{2kl}$, for $k = x_1$. We proceed with $X_3, \ldots, X_7$. In that order, the parent node has always been sampled \textit{before} the node itself.
4. (1 point) Convert this graph into an undirected graph. (Explain.)

**Solution: Moralisation:** replace each directed edge by an undirected edge, and marry the parents. For a directed tree, it is enough to remove the tip of each arrow, as nodes have either one or zero parent.

5. (1 point) Based on this graph, can we say that $X_3$ is independent from $X_5$? independent from $X_5$ conditional on $X_4$?

**Solution:** No, and yes. There is a path between $X_3$ and $X_5$ (so they are not independent), but this path is "blocked" as soon as we put our finger on $X_4$; hence they are conditionally independent, given $X_4$.

We now wish to perform probabilistic inference (i.e., computing marginal distributions and related tasks, as explained in the course). We start from the undirected graph. For the considered distribution, you may use the same notations as in the course for undirected graphs; i.e.,

$$p(x) = \frac{1}{Z} \prod_{i=1}^{n} \psi_i(x_i) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$$

without any connection (for now) with the previous directed graph.

6. (2 points) We have seen that probabilistic inference is simple to implement for graphs that are chains. Explain a simple way to transform the undirected graph into a chain. What is the complexity of the resulting algorithm?

**Solution:** We could merge nodes 4 and 6, and nodes 5 and 7; the pairs of variables $(X_4, X_6)$ and $(X_5, X_7)$ may be treated as categorical variables with $K^2$ values. Note that this is a simple application of the junction tree algorithm. Since the complexity (in $K$) of probabilistic inference is $O(K^3)$ in the standard case, here the complexity would be $O(K^4)$.

7. (3 points) The previous approach seems too expensive when $K$ is large. Instead let's go back to the undirected graph. Explain how you may use belief propagation to compute the marginal distribution of the central node. Re-derive the algorithm for this simple graph (i.e., give the corresponding expressions for all the involved quantities).

**Solution:**

First, we declare $X_4$ to be the root of the tree. This makes $X_2$, $X_4$ and $X_6$ the direct children of the root, and $X_1$, $X_5$ and $X_7$ the grand-children (and the terminal leaves).

Using the standard notations recalled above, we are trying to compute:
\[
p(x_3) = \sum_{x_1, x_2, x_4, \ldots, x_7} p(x) \\
= \frac{\psi_3(x_3)}{Z} \sum_{x_1} \psi_1(x_1) \sum_{x_2} \psi_{1,2}(x_1, x_2) \psi_2(x_2) \psi_{2,3}(x_2, x_3) \\
\times \sum_{x_5} \psi_5(x_5) \sum_{x_4} \psi_{4,5}(x_4, x_5) \psi_4(x_4) \psi_{3,4}(x_3, x_4) \\
\times \sum_{x_7} \psi_7(x_7) \sum_{x_6} \psi_{6,7}(x_6, x_7) \psi_6(x_6) \psi_{3,6}(x_3, x_6)
\]

(All the sums are for \( x_i = 1 \) to \( x_i = K \).)

By re-arranging the terms as we did above, we already see that the marginal distribution of \( x_3 \) is expressed as a product of three factors, corresponding to the three branches that connect at node \( X_3 \). In addition, we notice that we can compute recursively these factors, using message passing; for instance, for the first branch, define message

\[
\mu_{1 \rightarrow 2}(x_2) = \sum_{x_1} \psi_1(x_1) \psi_{1,2}(x_1, x_2)
\]

then this message costs \( \Theta(K) \) to compute, for each value of \( x_2 \) (hence \( K^2 \) in total), and the factor of the first branch equals

\[
\sum_{x_2} \mu_{1 \rightarrow 2}(x_1, x_2) \psi_2(x_2) \psi_{2,3}(x_2, x_3)
\]

which is simply the message \( \mu_{2 \rightarrow 3}(x_3) \), which, again, costs \( \Theta(K) \) to compute. Apply the same recipe for the two other branches.

We now go back to the initial (directed) graph.

8. (1 point) Explain how you may use the approach proposed in Question 6 to compute the marginal distribution of the central node; in particular, explain how to choose the functions \( \psi_i \) and \( \psi_{ij} \), based on the parameters defined in Question 2.

**Solution:** In the generic expression for distributions for undirected graphs (with functions \( \psi_i \) and so on); take \( Z = 1 \), \( \psi_1(x_1) = \pi_{x_1} \), and \( \psi_{i,j}(x_i, x_j) = A_{j,x_i,x_j} \) for any \((i, j)\) such that \( i \) is the parent of \( j \); all the other functions are set to constant \( \text{one} \).

Then we may apply the belief propagation algorithm to recover the marginal distribution of \( X_3 \).

9. (2 points) Find a simpler way to compute the marginal distribution of the directed graph.

**Solution:** Actually, for node \( X_3 \), the approach based on the undirected graph is not the fastest one; we may simply compute it by, essentially, passing messages from \( X_1 \) to \( X_3 \); or with symbols:

\[
P(X_3 = l) = \sum_{x_1} \pi_{x_1} \sum_{x_2} A_{2,x_1,x_2} A_{3,x_2,x_3}
\]
where this double sum may be computed using the same recursion as in message passing. (But note that we have only two messages to compute here.)

3 Noisy Ising

Consider a simple version of the Ising model:

\[ p(x) = \frac{1}{Z_{\alpha,\beta}} \exp \left\{ \alpha \sum_{i=1}^{n} x_i + \beta \sum_{(i,j) \in E} 1(x_i = x_j) \right\}, \]

where the \( x_i \) are either 0 or 1. (Recall that \((i,j) \in E\) means that there is an edge between nodes \( i \) and \( j \), and that \( 1(x_i = x_j) \) is one if \( x_i \) and \( x_j \) have the same “colour”, zero otherwise).

We do not observe \( x \) directly. Instead, we observe independent variables \( y_i \) which, conditional on \( x_i = l \) are distributed according to a Gaussian distribution \( N(\mu_l, 1) \). This type of model is often used in, e.g., medical imagery (as a way to automatically detect certain regions, e.g., bones).

We first treat the parameters \( \alpha, \beta, \mu_0, \mu_1 \) as fixed.

1. (2 points) Explain the trick we may use to adapt belief propagation to the problem of computing the distribution of a given \( x_i \), conditional on all the \( y_i \)'s. Is this algorithm exact? Explain.

\textbf{Solution:} We need to introduce \textbf{Dirac} potential functions \( \psi_{y_i}(y_i) = \delta(y_i, \bar{y}_i) \), where \( \bar{y}_i \) denotes the actual observation. Then we may apply belief propagation in the usual way.

This algorithm is not exact in this case, because the graph has \textbf{loops} (or, in other words, it is not a tree). Instead, we may iterate message passing until convergence, but the algorithm does not converge to the true solution.

2. (3 points) Explain how you may instead use MCMC to sample from the same conditional distribution. Compute any distribution you need to sample from, and explain why these distributions are easy to sample from.

\textbf{Solution:} We may use \textbf{Gibbs sampling}, a particular MCMC algorithm where we iteratively sample from each component \( x_i \), conditional on all the other variables. In this particular case, we find that:

\[ p(x_i|x_{-i}) \propto \exp \{ \alpha x_i + \beta \sum_j 1[x_i = x_j] + (y_i - \mu_{x_i})^2/2 \} \]

where the sum is over the \textbf{neighbours} of \( i \). Since \( x_i \) may take only two values, we can easily normalise (and sample from this distribution).

\textbf{Note:} the expression above is obtained by first writing down the joint distribution of all the \( x_i \)'s and the \( y_i \)'s, and then by dropping all the factors that do not depend on \( x_i \), as explained during the course.

We now wish to learn the parameters, based on data \( y = (y_1, \ldots, y_n) \). For simplicity, we assume first \( \alpha \) and \( \beta \) are fixed, and we want to learn \( \mu_0 \) and \( \mu_1 \).
3. (2 points) Derive the update equation of the EM algorithm for this model. How could you use Question 2 to actually implement such an EM algorithm? Elaborate and discuss.

Solution: The complete log-likelihood is (up to a constant that does not depend on the parameters):

$$\alpha \sum_i x_i + \beta \sum_{(i,j) \in E} 1(x_i = x_j) - (1/2) \sum_i (y_i - \mu_{x_i})^2$$

and its expectation with respect to $X|Y = y$, is:

$$cst - (1/2) \sum_i \left[ P(x_i = 0|y)(y_i - \mu_0)^2 + P(x_i = 1|y)(y_i - \mu_1)^2 \right]$$

so if we maximise with respect to $\mu_0$ (say), we get:

$$\mu_0 = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

where the weight is $w_i = P(x_i = 0|y)$ (which depends on the current value of the parameters). These probabilities may be computed in various ways; we could use belief propagation, as in Question 1, but this would be approximate. Or we can use Gibbs sampling, as in Question 2, but we then get some Monte Carlo error (and the algorithm is then pretty expensive, as we must run Gibbs sampling at each EM iteration!).

4. (2 points) Now assume that $\alpha$ and $\beta$ are also unknown. What problem arises in this case, if we try to implement the EM algorithm?

Solution: In this case, in the calculations above, we cannot get rid of

$$-\log Z_{\alpha,\beta}$$

the log of the normalising constant (partition function). We cannot easily maximise the corresponding function wrt $\alpha$ and $\beta$. 

Page 8