Problem 1 – Poisson distribution as exponential family
Let $\Omega = \{0, 1, \ldots n\ldots\}$ the set of non-negative integers. The Poisson distribution with parameter $\lambda > 0$ is defined by

$$P(n) = \frac{\lambda^n}{n!} e^{-\lambda}, \text{ for } n \in \Omega$$

(a) Find the natural parametrization of the Poisson family, $P_\theta(n) = e^{\theta n - \psi(\theta)}$, calculate $\theta$ as a function of $\lambda$, and $\psi(\theta)$.

Find the domain $\Theta$ of $\theta$.

[b – Not graded] Note that this is a family where $c(n) \neq 0$ and that omitting $c(n)$ results in another (exponential) family of distributions over the natural numbers.

[c. – Not graded] Find the expectation $E_\theta[n]$ by taking the derivative of $\psi(\theta)$, as well as the variance of $n$. Verify that they are equal to $\lambda$.

[d. Not graded] Find the Mean-Value parametrization of the Poisson distribution. Show that $\lambda$ is the mean value parameter, and infer the marginal polytope $M = \{\mu | E_\theta[n] = \mu, \theta \in \Theta\}$.

[e. – Not graded] Derive the ML estimation equations for $\theta$ and $\mu \equiv \lambda$. Note that for mean-value parametrization, the ML estimation equations are particularly simple.

f. – Single layer neural net with Poisson regression
Let now $\theta = \beta^T x$ with $x \in \mathbb{R}^d$ a vector of observed inputs. Assume that we observe an i.i.d. dataset $D = \{(x_1, y_1), \ldots (x_N, y_N)\}$. Write the expression of $l(\beta) = \frac{1}{N} \sum \ln P(y_i | \beta, x_i)$.

Calculate the expression of $\frac{\partial l}{\partial \beta_j}$.

Problem 2 – Fisher Information
The Fisher information $I(\theta)$ of a distribution of $X \in \Omega$, depending on parameter $\theta$ is defined as $I(\theta) = E_{P_\theta}[-\nabla^2 \ln P_\theta(X)]$. If $\theta$ is a $d$-dimensional vector, then $I(\theta)$ is a $d \times d$ symmetric matrix.

(a) Show that the Fisher information for an exponential family model

$$P_\theta(x) = e^{\theta^T t(x) - \psi(\theta)}$$

is equal to $\nabla^2 \psi(\theta)$. 

Read: Generalized Linear Models from Lecture 3
b. Assume for simplicity that $\psi$ and its derivatives are available in closed form. We observe data $x^{1:N}$ (i.i.d.) and want to obtain the ML estimate $\theta$ iteratively, by the Newton\(^1\) iteration.

Derive the Newton step for the ML estimation of $\theta$ as a function of the Fisher information and other statistical quantities related to $P_\theta$.

c. **KL divergence and Fisher information** Consider now any parametric family $P_\theta(x)$, $x \in \Omega$. Denote by $g_\theta(\theta') = KL(P_\theta||P_{\theta'})$. Calculate $\nabla g_\theta = \frac{\partial}{\partial \theta} g_\theta(\theta')$. Calculate $\nabla^2 g_\theta$ and show that at $\theta = \theta'$, $\nabla^2 g_\theta$ related to the Fisher information. (Note that if $\theta \in \mathbb{R}^k$, then $\nabla g \in \mathbb{R}^k$, $\nabla^2 g \in \mathbb{R}^{k \times k}$.)

[d. – Extra credit] Take the Taylor expansion of $g_\theta$ around $\theta' = \theta$ up to order two, and derive from it a simple approximation of the KL-divergence by means of the Fisher information.

**Problem 3 – Conditional distributions in the Ising model (only c,d graded)**

Here we will show that in an Ising model the conditional distribution of a single variable given the others has a simple expression.

Assume an Ising model over $V = \{1, 2, \ldots, n\}$ is defined by

$$\ln P_V = \sum_{i \in V} a_i x_i + \sum_{ij \in E} b_{ij} x_i x_j - \psi$$

(3)

In the above, $E$ is a given set of edges (i.e pairs of variables), $x_i \in \{\pm 1\}$, $i \in V$ and $a_i, i \in V, b_{ij}, ij \in E$ are real parameters. The normalization constant $\psi$ depends on $E$ and the parameters $a_i, b_{ij}$.

a. We will start with an example having only 5 variables.

For the graph above, consider the corresponding Ising model and calculate the expression of

$$\frac{P(x_1 = +1, x_2, x_3, x_4, x_5)}{P(x_1 = -1, x_2, x_3, x_4, x_5)}$$

(4)

as a function of $a_i, i \in V, b_{ij}, ij \in E$ and the variables $x_2, x_3, x_4, x_5$.

b. From the expression you found in a, derive an expression for the conditional probability

$$P(x_1 = +1|x_2, x_3, x_4, x_5)$$

Simplify the expression as much as possible.

c. Now for an arbitrary undirected graph with the Ising model (3), calculate the expression

$$\frac{P(x_1, x_2, \ldots, x_{i-1}, +1, x_{i+1}, \ldots, x_n)}{P(x_1, x_2, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n)}$$

(5)

\(^1\)The Newton step for minimizing a function $f(z)$ is given by $z^{k+1} = z^k - (\nabla^2 f(z^k))^{-1} \nabla f(z^k)$; see also STAT 535 notes and textbook.
as a function of $a_i, i \in V, b_{ij}, ij \in E$ and the variables $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots x_n$

d. From the expression you found in c, derive an expression for the conditional probability

$$P(x_i = +1 | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots x_n)$$

Simplify the expression as much as possible.
[Optional: where have you encountered this distribution before in your statistics studies?]

[e. – Not graded] Use the result in d. to conclude that the Markov blanket of any variable $X$ in an Ising model is represented by $n(X)$.

**Problem 4 – Marginals in Markov field**

We will look at two Ising models which differ in structure, but have “the same” edge potentials. Define the function $\phi(x, x')$ by

$$\phi(0, 0) = \phi(1, 1) = 1 \quad \phi(1, 0) = \phi(0, 1) = \theta \in (0, 1) \quad (6)$$

For a set of nodes $V$ and a set of edges $E$, define

$$P_{V,E}(x_V) = \frac{1}{Z_{V,E,\theta}} \prod_{ij \in E} \phi(x_i, x_j) \quad (7)$$

where it is assumed that the variable associated to each node $i \in V$ is $X_i \in \{0, 1\}$.

**A: 4 nodes  B: 9 nodes**

For the two graphs shown here

$$\begin{array}{cccc}
1 & 2 & -5 \\
3 & 4 & -6 \\
7 & 8 & -9
\end{array}$$

and for $\theta = 0.2$, do the following.

a. Compute numerically (by brute force) the normalization constants $Z^A, Z^B$.

b. Compute the marginal probability $P_{12}^{A,B}$ of variables $X_1, X_2$ under either model.

c. Compare these marginals, and compare them with $\phi$. Are they equal or proportional to one another?