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

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

(1)

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 \( l(\beta) = \frac{1}{N} \sum_i \ln P(y_i|\beta, x^i) \).

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

Problem 2 – Entropies, KL divergences, mutual informations

a. Calculate the entropy of a univariate normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Explain why the expression you obtain does not depend on \( \mu \).

[b. - Optional, extra credit] Calculate the entropy of a multivariate normal distribution with mean 0 and covariance matrix \( \Sigma \).

[c. Calculate the KL divergence between two univariate normal distributions \( N(\mu_1, \sigma_1^2) \), \( N(\mu_2, \sigma_2^2) \).
d. Let $X,Y \in \mathbb{R}$ be jointly normal with means 0. Find the expression of the mutual information $I(X \mid Y)$.

e. $X,Y \in \{0,1\}$ have joint distribution given by $P[X = 1, Y = 1] = P[X = 0, Y = 0] = p/2$, $P[X = 1, Y = 0] = P[X = 0, Y = 1] = q/2$, $p + q = 1$. Find the expression of the mutual information $I(X,Y)$ as a function of $p,q$. For what values of $p,q$ is $I(X,Y)$ maximized and minimized? What other information theoretic expression is represented by $I(X,Y)$?

f. Let $X \in \{0,1\}$, $Y \in \mathbb{R}$ be two random variables, with $P_X(1) = p$, $P_Y \mid X = N(X, \sigma^2)$. In other words, $P_Y$ is a mixture of two normal distributions.

For $p = 0.25$ and $p = 0.5$ calculate (numerically) the entropy $H(Y; \sigma)$ as a function of $\sigma$. Plot the values obtained for $\sigma \in [0.05, 1]$.

g. Explain how you can obtain the mutual information $I(X,Y)$ using the numerical values computed above. Plot $I(X,Y)$ as a function of $\sigma$ for $p = 0.25$ and $p = 0.5$, $\sigma \in [0.05, 1]$.

Explain in a sentence or two the behavior observed towards small $\sigma$ values. What is the value towards which $I(X,Y)$ converges when $\sigma \to 0$? (No rigorous proof required for this question.)

**Problem 3 – Fisher Information**

The Fisher information $I(\theta)$ of a distribution of $X \in \Omega$, depending on parameter $\theta$ is defined as $I(\theta) = E_{\mathcal{P}_\theta}[-\nabla^2 \ln P(\theta)]$. 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 $\nabla^2 \psi(\theta)$.

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 $\hat{\theta}$ iteratively, by the Newton 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.

---

1The 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.