Probabilistic Inference & Learning

Exercise Sheet #9

Exponential Families Reminder: Consider a random variable X taking values $x \in \mathbb{X} \subset \mathbb{R}^n$. A probability distribution for X with pdf of the form

$$p(x \mid w) = \exp\left(\phi(x)^\intercal w - \log Z(w)\right), \qquad \text{where} \qquad Z(w) := \int_{\mathbb{X}} \exp\left(\phi(x)^\intercal w\right)\, dx, \tag{1}$$

is called an **exponential family** of probability distributions. The function $\phi : \mathbb{X} \to \mathbb{R}^d$ is known as **the sufficient statistics** of p. The parameters $w \in \mathbb{R}^d$ are **the natural parameters** of p. The constant $Z(w) : \mathbb{R}^d \to \mathbb{R}$ is **the partition function**.
(For the following exercises, it is helpful to note that w and ϕ are only defined relative to each other, so if (ϕ, w) is a parametrization of an exponential family, then so is $(A\phi, A^{-1}w)$ for any invertible linear map A).

1. **Famous Exponential Families**: Many popular probability distributions are exponential families. Rewrite the following four distributions in the functional form of Eq. (1) and explicitly identify one possible choice for sufficient statistics ϕ, the natural parameters w, and the partition function $Z(w)$:

   (a) the **Bernoulli** distribution (with $f \in [0, 1]$) *10 points*

   $$p(x \mid f) = f^x \cdot (1 - f)^{1-x} \text{ for } x \in \{0; 1\} \tag{2}$$

   (b) the **Gamma** distribution (with $\alpha, \beta \in \mathbb{R}_+$) *10 points*

   $$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} =: \mathcal{G}(x; \alpha, \beta) \tag{3}$$

   (c) the **Dirichlet** distribution (with $\alpha \in \mathbb{R}^n_+$) *10 points*

   $$p(x \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \text{ for } x \in \left\{ x \in \mathbb{R}^n \middle| \sum_i x_i = 1 \right\} \tag{4}$$

   (d) the **multivariate Gaussian** distribution (with $\mu \in \mathbb{R}^n$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{n \times n}$) *10 points*

   $$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^\intercal \Sigma^{-1} (x - \mu) \right) \text{ for } x \in \mathbb{R}^n \tag{5}$$

2. **Maximum Likelihood Inference** It was shown in the lecture that for draws $x_i \in \mathbb{X}$, $i = 1, \ldots, m$, the *maximum likelihood estimate* $w_{\text{ML}} = \arg\max_w \prod_{i=1}^m p(x_i \mid w)$ of the natural parameters w can be computed by computing the empirical mean estimate of the sufficient statistics and solving the equation

   $$\frac{1}{m} \sum_i \phi(x_i) = \nabla_w \log Z(w) \tag{6}$$

   for w. Find the value of w_{ML} in this way for the two exponential families listed in 1.(a) and 1.(d), and express it in terms of the standard (as opposed to natural) parameters listed in respective exercise. (For 1.(b) and 1.(c) this equation has no analytic solution, although it can be solved efficiently using numerical methods) *30 points*

3. **Conjugate Prior Inference** The following is the re-construction of a famous result by William Sealy Gosset (1876–1937), a statistician (a student of Pearson) working for the Guinness brewery in Dublin who, in his work, had to estimate the likely error of the estimate of the mean of a Gaussian distribution constructed from a small sample of assays (he would later go on to become the Chief Brewer). His employer prohibited him from publishing his work under his own name, so he used the pseydonym *Student*. This exercise is actually only a part of his main result; the full story (conjugate prior inference on mean and variance of a Gaussian) is a bit too long for this exercise sheet.

   Assume we are given m independent draws $x_i \in \mathbb{R}, i = 1, \dots, m$ from a *scalar* Gaussian distribution. Assume that we know the distribution's mean μ, but not the variance σ^2:

$$p(x_i \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{7}$$

   (a) By re-writing Equation (7) in the form of (1), show that this likelihood is an exponential family on x with the natural parameter $w = \sigma^{-2} \in \mathbb{R}_+$. What are the sufficient statistics, and the partition function? *10 points*

   (b) Show that the **Gamma-distribution** $\mathcal{G}(w; \alpha, \beta)$ (cf. Eq. (3)) is a conjugate prior on w for Eq. (7). That is, show that when using (3) as the prior on w, one can write the posterior on w from the m observations as

$$p(w \mid x_1, \dots, x_m) = \frac{\mathcal{G}(w; \alpha, \beta) \prod_{i=1}^m p(x_i \mid w = \sigma^{-2})}{p(x_1, \dots, x_m)} = \mathcal{G}(w; \alpha_m, \beta_m) \tag{8}$$

   with new parameters α_m, β_m. What are these new parameters? *10 points*

   (c) We already know from 1.(b) that the Gamma distribution is also an exponential family. Its partition function is given through the third Eulerian integral (the Gamma function Γ) as

$$Z_{\mathcal{G}}(\alpha, \beta) = \int_0^\infty w^{\alpha-1} \exp(-\beta w)\, dw = \beta^{-\alpha}\Gamma(\alpha). \tag{9}$$

   Use this integral to show that when using (3) as the conjugate prior, and given the m observations (i.e. given the posterior $p(w \mid x_1, \dots, x_m) = \mathcal{G}(w; \alpha_m, \beta_m)$), the *predictive* distribution for the *next* observation x_{m+1} is given by the **Student-t** distribution *10 points*

$$p(x_{m+1} \mid \alpha_m, \beta_m) = \int_0^\infty p(x_{m+1} \mid w)\mathcal{G}(w; \alpha_m, \beta_m)\, dw \tag{10}$$

$$= \frac{\Gamma(\alpha_m + 1/2)}{\Gamma(\alpha_m)} \frac{1}{\sqrt{2\pi\beta_m}} \left(1 + \frac{(x - \mu)^2}{2\beta_m}\right)^{-\alpha_m - 1/2}. \tag{11}$$