Deadline: Th. April 25, 10:00 am Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz. Alternatively upload a Jupyter notebook (`.ipynb`) or a `.pdf` file via LearnWeb.

Definition 1. A probability density $p(x;\theta)$ with parameters θ belongs to the exponential family, if it can be expressed in the form

$$p(x;\theta) = h(x)e^{\langle \eta(\theta) | \Phi(x) \rangle - A(\theta)} \tag{1}$$

for some functions h, η, Φ, A

1 Exponential Family (10 points)

A. [3p] Show that the Gamma distribution $\Gamma(\alpha, \beta)$ defined by

$$p(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0 \tag{2}$$

belongs to the exponential family.

B. [3p] Show that the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ defined by

$$p(x;\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \quad \mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)} \tag{3}$$

belongs to the exponential family.

C. [4p] Show that the uniform distribution $\mathbb{1}_{[a,b]}$ defined by

$$p(x;a,b) = \begin{cases} \frac{1}{b-a} : & x \in [a,b] \\ 0 : & \text{else} \end{cases} \tag{4}$$

does not belong to the exponential family. Hint: For a continuous function f, its so called *support* is defined as $\operatorname{supp} f = \{x \mid f(x) \neq 0\}$. What is $\operatorname{supp} \frac{h(x)}{Z(\theta)} e^{\phi(x)\eta(\theta)}$? What is $\operatorname{supp} \mathbb{1}_{[a,b]}$?

2 Generalized Linear Models (10 points)

A. [3p] What are the 3 components of a GLM? What is their purpose?

B. [7p] Consider a dataset (X, Y), with x, y scalar, such that a linear change in x causes a percentage change in y, suggesting a model of the form $y \propto \exp(\beta^T x)$. Moreover it also appears that the Y data is always non-negative. We are faced with two different modeling approaches:

- Approach A: Transform the initial data from (X, Y) to $(X, \log(Y))$. Then apply a standard linear regression ($\rightsquigarrow \mathbb{E}[\log y] = \beta^T x$)
- Approach B: Construct a GLM satisfying $\log \mathbb{E}[y] = \beta^T x$.

What is problematic about approach A, and how is handled better by approach B? What could be an appropriate member of the exponential family and link function for the second approach?

3 (Bonus) Maximum Entropy Distributions (10 points)

In this problem we want to show that the Normal Distribution $\mathcal{N}(\mu, \sigma^2)$ is the maximum entropy distribution on $\mathbb{R}$ with the side conditions $\mathbb{E}[x] = \mu$ and $\mathbb{V}[x] = \sigma^2$ To show this we are faced with a variational optimization problem: We want to find the function f which maximizes the entropy term

$$H(f) = -\int_{-\infty}^{+\infty} f(x) \log f(x) \mathrm{d}x \tag{5}$$

satisfying the side conditions (using second moment instead of variance)

$$m_0(f) = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1, \quad m_1(f) = \int_{-\infty}^{+\infty} x f(x)\mathrm{d}x = \mu, \quad m_2(f) = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x = \sigma^2 + \mu^2$$

To transform the constrained problem into an unconstrained one, consider the Lagrangian

$$\mathscr{L}(f, \lambda) = H(f) + \lambda_0(m_0(f) - 1) + \lambda_1(m_1(f) - \mu) + \lambda_2(m_2(f) - (\mu^2 + \sigma^2)) \tag{6}$$

A necessary condition for a local extrema is that the derivative of the Lagrangian is zero. Note that $\mathscr{L}$ depends on a function! ($\rightsquigarrow$ functional derivative). Equivalently, one can show that all partial derivatives are zero. To proceed we will need the following theorem:

Theorem (Fundamental lemma of calculus of variations)**.** $\boxed{\textit{If } \int_a^b f(x)g(x)\mathrm{d}x = 0 \textit{ for all } g, \textit{ then } f = 0}$

The idea now is to consider an arbitrary function g and compute the partial derivative

$$\frac{\partial}{\partial \epsilon} \mathscr{L}(f + \epsilon g, \lambda)\bigg|_{\epsilon=0} = 0 \tag{7}$$

Here $\frac{\partial}{\partial \epsilon} \mathscr{L}(f + \epsilon g, \lambda)\big|_{\epsilon=0}$ is the derivative of the function $\epsilon \to \mathscr{L}(f + \epsilon g, \lambda)$, evaluated at $\epsilon = 0$. Note that this function depends only on a scalar! For example, $\frac{\partial}{\partial \epsilon} H(f + \epsilon g)$ is equal to:

$$-\frac{\partial}{\partial \epsilon} \int_{-\infty}^{+\infty} (f + \epsilon g) \log(f + \epsilon g)\mathrm{d}x = -\int_{-\infty}^{+\infty} \frac{\partial}{\partial \epsilon}(f + \epsilon g) \log(f + \epsilon g)\mathrm{d}x = -\int_{-\infty}^{+\infty} g + g \log(f + \epsilon g)\mathrm{d}x$$

hence $\frac{\partial}{\partial \epsilon} H(f + \epsilon g)\big|_{\epsilon=0} = \int_{-\infty}^{+\infty} g(1 + \log(f))\mathrm{d}x$

1. Show that $\frac{\partial}{\partial \epsilon} \mathscr{L}(f + \epsilon g, \lambda)\big|_{\epsilon=0} = \displaystyle\int_{-\infty}^{+\infty} (1 + \log f(x) + \lambda_0 + \lambda_1 x + \lambda_2 x^2) g(x)\mathrm{d}x$

2. Apply the fundamental lemma. Conclude that f is of the form $f(x) = c e^{a(x-b)^2}$ by substituting λ for some appropriate (a, b, c)

3. Use the constraints and the formulas below to determine a, b, c. (The result should be the Normal distribution!)

$$\int_{-\infty}^{+\infty} e^{-a(x-b)^2}\mathrm{d}x = \sqrt{\tfrac{\pi}{a}}, \quad \int_{-\infty}^{+\infty} x e^{-a(x-b)^2}\mathrm{d}x = \sqrt{\tfrac{\pi}{a}}b, \quad \int_{-\infty}^{+\infty} x^2 e^{-a(x-b)^2}\mathrm{d}x = \sqrt{\tfrac{\pi}{a}}(b^2 + \frac{1}{2a})$$