Exercise 1: Connection between MLE and ERM

Imagine you work at a car dealer and are tasked with predicting the monthly number of cars that will be sold within the next year. You decide to address this challenge in a data-driven manner and develop a model that predicts the number of cars from data regarding vehicles' properties from sold cars of previous years, current competitor and market data.

a) Let x_1 and x_2 measure the number of sold cars of the previous month and of the previous year, respectively. Both features and target are numeric and discrete. You choose to use a generalized linear model (GLM) for this task. For this, you assume the targets to be conditionally independent given the features, i.e., $y^{(i)}|\mathbf{x}^{(i)} \perp y^{(j)}|\mathbf{x}^{(j)}$ for all $i, j \in \{1, 2, \dots, n\}, i \neq j$, with sample size n.

- Argue which of the following distributions from the one-parametric exponential family is most suitable for the underlying use case: normal, Bernoulli, gamma or Poisson.
- Write down the probability distribution of the chosen distribution depending on $\boldsymbol{\theta}$ assuming a log link function.

b) State the hypothesis space for the corresponding model class. For this, assume the parameter vector $\boldsymbol{\theta}$ to include the intercept coefficient.

c) Which parameters need to be learned? Define the corresponding parameter space Θ.

d) In classical statistics, you would estimate the parameters via maximum likelihood estimation (MLE).

Describe how you can make use of the likelihood in empirical risk minimization (ERM) and write down the likelihood as well as the resulting empirical risk.