Solution 1:

a) The normal and gamma distribution assume a continuous target variable, however, we do have a discrete target variable. Since the target variable (number of cars) represents a count variable, Bernoulli (taking only the values 0 and 1) is not a suitable choice in this context. It follows that the only reasonable choice of the given distributions is the Poisson distribution. The Poisson distribution depends on the parameter λ, where the expected value is given by $E(Y|\mathbf{x}) = \lambda(\mathbf{x})$. The log link function is given by $\log(\lambda(\mathbf{x})) = \boldsymbol{\theta}^T\mathbf{x}$. Following from that the probability function is given by

$$P(Y = y) = \frac{\exp\left(-\lambda(\mathbf{x})\right) \cdot (\lambda(\mathbf{x}))^y}{y!} = \frac{\exp\left(-\exp\left(\boldsymbol{\theta}^T\mathbf{x}\right)\right) \cdot (\exp\boldsymbol{\theta}^T\mathbf{x})^y}{y!}$$

for $y \in \mathbb{N}_0$.

b) We can write the hypothesis space as:

$$\mathcal{H} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T\mathbf{x}) \mid \boldsymbol{\theta} \in \mathbb{R}^3\} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \mid (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3\}.$$

Note the **slight abuse of notation** here: in the lecture, we first define $\boldsymbol{\theta}$ to only consist of the feature coefficients, with $\mathbf{x}$ likewise being the plain feature vector. For the sake of simplicity, however, it is more convenient to append the intercept coefficient to the vector of feature coefficients. This does not change our model formulation, but we have to keep in mind that it implicitly entails adding an element 1 at the first position of each feature vector, i.e., $\mathbf{x}^{(i)} := (1, x_1, x_2)^{(i)} \in \{1\} \cup \mathcal{X}$, constituting the familiar column of ones in the design matrix $\mathbf{X}$.

c) The parameter space is included in the definition of the hypothesis space and in this case given by $\Theta = \mathbb{R}^3$.

d) The likelihood for the Poisson distribution is defined by:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} \frac{\exp\left(-\exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)\right) \cdot \left(\exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)\right)^{y^{(i)}}}{y^{(i)}!}$$

The first thing to note is that both MLE and ERM are **optimization problems**, and both should lead us to the same optimum. Their opposite signs are not a problem: maximizing the likelihood is equivalent to minimizing the negative likelihood. Also, both are defined pointwise. The last thing to fix is therefore the product introduced by the independence assumption in the joint likelihood of all observations (recall that we use a *summed* loss in ERM), for which the logarithm is a natural remedy. We can thus simply use the **negative log-likelihood (NLL)** as our loss function (and indeed, many known loss functions can be shown to correspond to certain model likelihoods).

Let's put these reflections to practice:

$$\begin{aligned}
L_{NLL}\left(y^{(i)}, f\left(\mathbf{x}^{(i)}|\boldsymbol{\theta}\right)\right) &= -\log\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\
&= -\ell(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\
&= -\log\frac{\exp\left(-\exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)\right) \cdot \left(\exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)\right)^{y^{(i)}}}{y^{(i)}!} \\
&= \exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right) - y^{(i)}(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) + \log(y^{(i)}!)
\end{aligned}$$

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} -\ell(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\ &= \sum_{i=1}^{n} L_{NLL}\left(y^{(i)}, f\left(\mathbf{x}^{(i)}|\boldsymbol{\theta}\right)\right) \\ &= \sum_{i=1}^{n} \exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right) - y^{(i)}(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) + \log(y^{(i)}!) \\ &\propto \sum_{i=1}^{n} \exp\left(\boldsymbol{\theta}^T\mathbf{x}^{(i)}\right) - y^{(i)}(\boldsymbol{\theta}^T\mathbf{x}^{(i)})\end{aligned}$$

As we are only interested in the feature coefficients here, we neglect all irrelevant terms that do not depend on $\boldsymbol{\theta}$ as they have no effect on the solution (i.e., the $\arg\min$ of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$). This is what the proportional sign $\propto$, often used in contexts of optimization and Bayesian statistics, means: we keep only expressions impacted by our parameter of interest because they suffice to yield the intended results or show some property of interest.