Problem 1 (Linear Regression and Ridge Regression):

Let \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) where \( x_i \in \mathbb{R}^d \) and \( y_i \in \mathbb{R} \). The goal in linear regression is to find parameters \( w \in \mathbb{R}^d \) such that \( \forall i : y_i \approx w^T x_i \).

In the lecture we considered the \( \text{least-squares} \) optimization problem

\[
\arg\min_w \hat{R}(w) = \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2
\]

and showed that under some assumptions on \( D \) there exists a unique closed form solution

\[
w^* = (X^T X)^{-1} X^T y,
\]

where \( X \in \mathbb{R}^{n \times d} \) is a \( n \times d \) matrix with the \( x_i \) as rows and \( y \in \mathbb{R}^n \) is a vector consisting of the scalars \( y_i \).

(a) Show for \( n < d \) that (1) does not admit a unique solution and that \( w^* \) is ill-defined. Explain why in such a case we cannot uniquely identify \( w^* \).

(b) Consider the case \( n \geq d \). Under what assumptions on \( X \) does (1) admit a unique solution \( w^* \)? Give an example with \( n = 3 \) and \( d = 2 \) where these assumptions do not hold.

The \textit{ridge regression} optimization problem with parameter \( \lambda > 0 \) is given by

\[
\arg\min_w \hat{R}_{\text{Ridge}}(w) = \arg\min_w \left[ \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda w^T w \right].
\]

(c) Show that \( \hat{R}_{\text{Ridge}}(w) \) is convex with regards to \( w \) for the case \( d = 1 \).

(d) Derive the closed form solution \( w^*_{\text{Ridge}} = (X^T X + \lambda I_d)^{-1} X^T y \) to (2) where \( I_d \) denotes the identity matrix of size \( d \times d \).

(e) Show that (2) admits the unique solution \( w^*_{\text{Ridge}} \) for any matrix \( X \). Show that this even holds for the cases in (a) and (b) where (1) does not admit a unique solution \( w^* \).

(f) What is the role of the term \( \lambda w^T w \) in \( \hat{R}_{\text{Ridge}}(w) \)? What happens to \( w^*_{\text{Ridge}} \) as \( \lambda \to 0 \) and \( \lambda \to \infty \)?

---

1 Without loss of generality, we assume that both \( x_i \) and \( y_i \) are centered, i.e. they have zero empirical mean. Hence we can neglect the otherwise necessary bias term \( b \).
Problem 2 (Normal Random Variables):
Let $X$ be a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\tau^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \tau^2)$. Recall that the probability density of $X$ is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-\mu)^2}{2\tau^2}}, \quad -\infty < x < \infty.$$ 

Furthermore, the random variable $Y$ given $X = x$ is normally distributed with mean $x$ and variance $\sigma^2$, i.e. $Y|_{X=x} \sim \mathcal{N}(x, \sigma^2)$.

(a) Derive the marginal distribution of $Y$.

(b) Use Bayes’ theorem to derive the conditional distribution of $X$ given $Y = y$.

Hint: For both tasks derive the density up to a constant factor and use this to identify the distribution.

Problem 3 (Bivariate Normal Random Variables):
Let $X$ be a bivariate Normal random variable (taking on values in $\mathbb{R}^2$) with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$. The density of $X$ is then given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

Find the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$. 

2