This exam consists of 4 questions, worth 45 points in total. Please write your name and student number on every sheet of your solutions.

**Question 1 [12 points]**
We draw a training dataset \((X_1, Y_1), \ldots, (X_m, Y_m)\) from a data generating distribution on \(\mathbb{R}^d \times \{-1, 1\}\). Suppose that \((X_1, Y_1), \ldots, (X_m, Y_m)\) are independent and identically distributed with \((X, Y)\). Assume, moreover, that there are some \(\bar{w} \in \mathbb{R}^d, \bar{b} \in \mathbb{R}\) and \(\frac{1}{2} < p < 1\) such that

\[
\mathbb{P}(Y = \text{sign}(\langle \bar{w}, X \rangle + \bar{b})) = p = 1 - \mathbb{P}(Y \neq \text{sign}(\langle \bar{w}, X \rangle + \bar{b})).
\]

(a) Give an interpretation (in words) for this model for data generation, in particular give an interpretation for \(p\).

(b) Show that the conditional probability mass function of \(Y\) given \(X\) is given by

\[
p(y|X) = p \left( \frac{1 - p}{p} \right)^{1\{\text{sign}(\langle \bar{w}, X \rangle + \bar{b}) \neq y\}}, \quad y \in \{-1, 1\}.
\]

(c) Show that \((w, b)\) is a maximum likelihood estimator of the pair \((\bar{w}, \bar{b})\) if and only if it is a solution of the optimization problem

\[
\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} 1\{\text{sign}(\langle w, X_i \rangle + b) \neq Y_i\},
\]

(d) (bonus question for 3 points) Is the statement in (c) still correct if \(p = 1\)?
(b) Show that (2) is equivalent to
\[
\min_{\alpha \in \mathbb{R}^m} \|y - G\alpha\|^2_2 + \lambda \alpha^T G\alpha,
\]
where \( G := [K(x_i, x_j)]_{i,j=1}^m \in \mathbb{R}^{m \times m} \) is the Gram matrix.

(c) Show that (3) is a convex optimization problem.

(d) Determine the set of all solutions of (3).

Question 3 [9 points]
Suppose that we draw a set of independent data points such that each point is identically distributed with \((X, Y)\), where \(X\) is uniformly distributed on \([-1, 1]\), \(\varepsilon\) is a standard normally distributed noise term, and \(Y = \frac{25}{2} X^3 + X^2 - \frac{1}{7} X + 2 + \varepsilon\). We split the data randomly into a training set (60%) and a validation set (40%). Consider the following hypothesis classes:

(1) \( f : \mathbb{R} \to \mathbb{R}, f(x) = a_0 + a_1 x, a_0, a_1 \in \mathbb{R} \)
(2) \( g : \mathbb{R} \to \mathbb{R}, g(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3, b_0, \ldots, b_3 \in \mathbb{R} \)
(3) \( h : \mathbb{R} \to \mathbb{R}, h(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_5 x^5 + c_6 x^6, c_0, \ldots, c_6 \in \mathbb{R} \)

For each class and every \( m \in \{1, \ldots, 15\} \) we train the model by minimizing the mean squared error (MSE) over the first \( m \) samples of the training data. Afterwards, we compute the mean squared error on the validation data. The result is visualized in the figure below. Match each class to the correct validation error curve and explain your reasoning carefully.

![Validation Error](image)

Question 4 [9 points]
Let \( K \subset \mathbb{R}^d \) be closed, non-empty, convex, and a cone, i.e., \( tx \in K \) whenever \( x \in K \) and \( t > 0 \). The polar cone of \( K \) is the set
\[
K^* = \{ y \in \mathbb{R}^d : \langle y, x \rangle \leq 0 \text{ for all } x \in K \}
\]
and the bipolar cone of \( K \) is \( K^{**} := (K^*)^* \).
(a) Show that $K \subset K^{**}$.

(b) Prove that $K^{**} \subset K$.

*Hint:* What would happen if $x \in K^{**}$, but $x \notin K$?