QUESTION 1

The dimensionality of a polynomial transform Φ_Q of order $Q = 10$ (excluding the additional coordinate $z_0 = 1$) is given by

$$\tilde{d} = \frac{Q(Q+3)}{2} = \frac{10 \cdot 13}{2} = 65. \qquad (1)$$

See the textbook p. 104 for reference.
The answer is therefore *e*: *None of the above*.

QUESTION 2

We consider the different options.
a *False.* In the singleton hypothesis set, g must be $\bar{g}$, so $\bar{g} \in \mathcal{H}$.
b *False.* If $g^{\mathcal{D}}$ are all real valued constants, their average $\bar{g}$ must be so as well, so $\bar{g} \in \mathcal{H}$.
c *False.* Averaging the weights of different linear regression models will always lead to another linear regression model.
d *True?* As it is unlikely that the answer to this question is "None of the above" again, it must be the logistic regression :)

QUESTION 3

In general, we have overfitting if – when transitioning from one hypothesis to another – the in-sample error E_{in} decreases while the out-of-sample error E_{out} increases.

a *True.* E_{in} has to decrease to meet the above definition.
b *True.* E_{out} has to increase to meet the above definition.
c *True.* Decreasing E_{in} and increasing E_{out} leads to a difference in $(E_{out} - E_{in})$.
d *False.* A change in $(E_{out} - E_{in})$ occurs most of the time, e.g. when E_{out} does not improve as fast as E_{in} during training (see lecture 11, slide 4).
e *True.* Overfitting is a relative term that is always based on the comparison between two or more hypotheses.
The correct answer is *d*.

QUESTION 4

We consider the different options.

a   *False.* Deterministic and stochastic noise can occur simultaneously.

b   *False.* Deterministic noise is the difference between the best fit hypothesis $h\star$ and the target function f. For different hypothesis sets, the deterministic noise can be different.

c   *False.* As in *b*, deterministic noise depends on the difference between $h\star$ and f, so it also depends on f.

d   *True.* Stochastic noise stems from deviations of the data from the target function f. These can be artifacts from data sampling or a result of a probability distribution in the data. It does not have anything to do with the hypothesis set.

e   *False.* The target distribution may directly affect stochastic noise as is produces data points that deviate from the expected value of the target function with a certain probability.

The correct answer is *d*.

QUESTION 5

The Tikhonov matrix Γ is an identity matrix scaled by a vector α. This can be used to regularize different weights (corresponding to the polynomial terms of different order) to individual degrees. From the textbook p. 130 we know that for the weight decay regularizer, the condition $\mathbf{w}_{\text{lin}}^T \mathbf{w}_{\text{lin}} \leq C$ leads to $\mathbf{w}_{\text{lin}} = \mathbf{w}_{\text{reg}}$. Now, one valid Tikhonov matrix to use would be $\Gamma = \mathbf{0}$ sow we'll get the weight decay regularizer. Therefore, the same condition must apply and $\mathbf{w}_{\text{lin}} = \mathbf{w}_{\text{reg}}$.

The correct answer is *a*.

QUESTION 6

The correct answer is *b*.

QUESTIONS 7–10

See the file `finalExam-question7-10.py` for the code. The correct answers are *d*, *b*, *e* and *a*.

See the file `finalExam-question11-12.py` for the code. From looking the the points in $\mathcal{Z}$-space, we can determine the support vectors and deduce the classifier line and margins from these. A plot of the data set including the classifier and margin lines is given in figure 1.

We can now deduce w_1, w_2 and b from the graph.

○ The classifier line is vertical, so the z_2 component does not have an influence and therefore $w_2 = 0$.

○ The z_1-components of the support vectors are either 0 or 1, therefore the vertical classifier line passes through $z_1 = 0.5$.

○ As all points on the decision boundary have $z_1 = 0.5$, we get $\mathbf{w}^T\mathbf{z} + b = w_1 \cdot 0.5 + b = 0$ when using $b = -0.5$ and $w_1 = 1$.

So we get $(w_1, w_2, b) = (1, 0, -0.5)$ which corresponds to solution c.

QUESTION 12

See the file `finalExam-question11-12.py` for the code. See figure 1 for the classifier boundary and support vectors in $\mathcal{X}$-space. The correct answer is c.
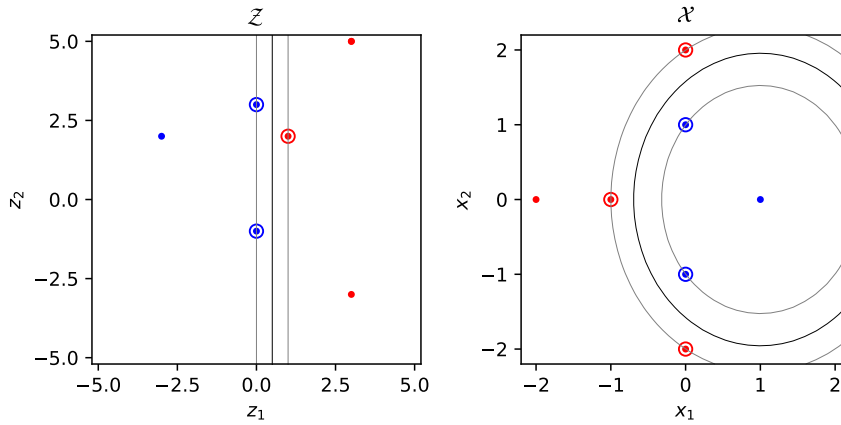

Fig. 1: Left: data set in $\mathcal{Z}$-space with support vectors, classifier line and margins marked. Right: data set in $\mathcal{X}$-space.

QUESTION 13–18

See the file `finalExam-question13-18.py` for the code. The correct answers are a, e, d, d, c and a.

The correct answer is *b*. This was found by an educated guess, so I'm looking forward to the solution.

Figure 2 shows different scenarios for a constant, real-valued hypothesis. We evaluate them to see how the error of the aggregate behaves. To answer the questions we apply the error measures specified in table 1.
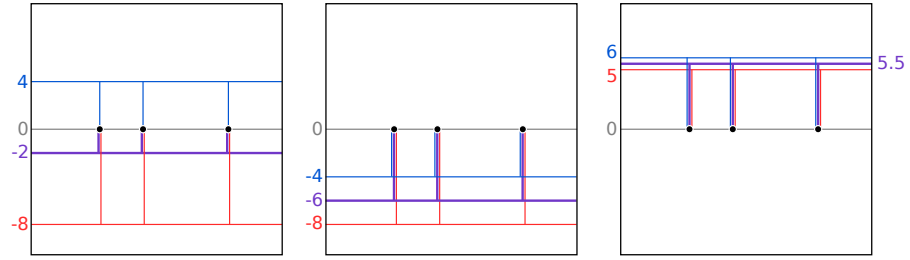

Fig. 2: Three different scenarios for the constant, real-valued hypothesis g_1, g_2 and the aggregate hypothesis g_{agg}. The training data is represented by the three dots (•)

Tab. 1: Error measures for the different choices *a-e*.

Error measure	Case 1	Case 2	Case 3
$E_{out}(g_1)$	16	16	36
$E_{out}(g_2)$	64	64	25
$mean(E_{out}(g_1), E_{out}(g_2))$	40	40	30.5
$min(E_{out}(g_1), E_{out}(g_2))$	16	16	25
$max(E_{out}(g_1), E_{out}(g_2))$	64	64	36
$E_{out}(g_{agg})$	4	36	30.25

We can now evaluate the answer choices with respect to these measures.

a $E_{out}(g_{agg}) \leq E_{out}(g_1)$: False by case 2

b $E_{out}(g_{agg}) \leq min(E_{out}(g_1), E_{out}(g_2))$: *False* by case 2

c $E_{out}(g_{agg}) \leq min(E_{out}(g_1), E_{out}(g_2))$: *True* for all cases

d $E_{out}(g_{agg}) \leq max(E_{out}(g_1), E_{out}(g_2)) \wedge E_{out}(g_{agg}) \geq min(E_{out}(g_1), E_{out}(g_2))$: *False* by case 1

The correct answer is *c*.