

[Return to Edit Assessment](#)

Name: _____

Score: _____ / _____

[Worksheet 01 - estimator theory](#)[Part 1](#)

Please state the names of all the students you worked with on this assignment:

*Answer Point Value: 0.0 points**Model Short Answer: -----*[Part 2: Characterization of the learning problem](#)

We are given the observation pairs $(\mathbf{x}_t, y_t)_{t=1, \dots, N}$ and we want to solve the following regression problem:

$$\min_{\theta} \sum_{t=1}^N (y_t - f(\mathbf{x}_t; \theta))^2$$

Characterize this learning problem.

- ☒ A. **Supervised**
- ☐ B. **Semi-supervised**
- ☐ C. **Unsupervised**
- ☐ D. **Generative**

*Answer Point Value: 1.0 points**Answer Key: A*

Which of the following aspects apply to the regression problem above?

- ☐ A. It minimizes the mean square error. (correct)
- ☐ B. It is Tschhebyscheff regression.
- ☐ C. It maximizes the likelihood of y_t having been emitted from $f(x_t; \theta)$ plus a Gaussian error. (correct)

Part 3: Modeling and numerics

Suppose we observe data that has been generated by a function of the form

$$y_i = a + bx_i + cx_i^2 + dx_i^4 + \epsilon$$

where ϵ is an iid normally distributed measurement error.

All coefficients are nonzero. We want to construct a linear least squares regression estimator:

$$\min_{\mathbf{w}} \sum_t \left(y_t - \sum_j w_j \phi_j(\mathbf{x}_t) \right)^2$$

Which of the following feature spaces can achieve a zero bias estimator?

- ☐ A. $\phi = (a, b, c, d)$
- ☐ B. $\phi = (x, x^2, x^4)$
- ☐ C. $\phi = (1, x, x^2, x^3, x^4)$
- ☐ D. None of these

Answer Point Value: 1.0 points

Answer Key: C

We use linear least squares (LLS) to fit a given dataset with the feature space $\phi = (x, x^2, \cos(x), \sin(x - \frac{\pi}{2}))$ (cos and sin are taken in radians).

What can be said about the regression result?

- ☐ A. The optimal LLS solution will have the form $\mathbf{w} = (a, b, 0, 0)$ where a and b depend on the data.
- ☐ B. The optimal LLS solution will have the form $\mathbf{w} = (a, b, c, -c)$ where a, b and c depend on the data.
- ☐ C. The feature correlation matrix $\mathbf{X}^\top \mathbf{X}$ is invertible.
- ☐ D. The optimal $L2$ -regularized result can be found with Ridge regression with nonzero λ parameter.

Answer Point Value: 1.0 points

Answer Key: D

Part 4: Hyperparameter selection

We fit a given dataset using Ridge regression with a polynomial model of the general form $f(x) = w_1 + w_2x + w_3x^2 + \dots w_nx^{n-1}$. We want to determine the maximum polynomial order with hyperparameter optimization.

Count the number of parameters and hyperparameters

- ☐ A. 1 parameter, n hyperparameters
- ☐ B. 2 parameters, n hyperparameters
- ☐ C. n parameters, 1 hyperparameters
- ☐ D. n parameters, 2 hyperparameters
- ☐ E. n parameters, n hyperparameters

Answer Point Value: 1.0 points

Answer Key: D

Suppose the observation data $(\mathbf{x}_t, y_t)_{t=1, \dots, N}$ has not been generated from a function that can be represented by a finite order polynomial, but we still want to approximate the function with a polynomial model. After 1000 datapoints, we conduct hyperparameter optimization and conclude that polynomial order n is optimal. Now consider we instead observe 100,000 datapoints coming from the same distribution.

Which polynomial order will now likely be optimal?

- ☐ A. $< n$
- ☐ B. n
- ☐ C. $> n$

Answer Point Value: 1.0 points

Answer Key: C

We have training data \mathbf{X}^{train} and validation data \mathbf{X}^{val} . We suppose that all data points from both datasets have been sampled independently and identically distributed from our data source. Suppose we train a model with parameters θ and hyperparameters λ . We have an error function E that serves as a loss function.

Now we want to construct an optimal estimator that minimizes the out-of-sample error E_{out} . What is the best strategy?

- ☐ A. Find an unbiased estimator, and then minimize the variance within unbiased estimators.
- ☐ B. Minimize the bias in the training error and minimize the variance in the validation error.
- ☐ C. Determine parameters by minimizing the loss function on the training set, determine hyperparameters by minimizing the loss function on the validation set.

Answer Point Value: 1.0 points

Answer Key: C

We have recorded a dataset \mathbf{X} . We are confident that we have enough data to have a representative sample, but we are unsure if two subsequently generated datapoints are statistically independent of another. Suppose we want to train a model with fixed hyperparameter settings and get a least biased estimate of the error on data not used for the training.

Which of the following strategies is best for this purpose?

- ☐ A. Use the first 80% of the data for training and the last 20% to compute the test error.
- ☐ B. Shuffle the data, i.e. randomly reorder the time points, and perform A.
- ☐ C. Perform five-fold cross-validation and use the mean validation error as test error.
- ☐ D. Repeat B 100 times and use the mean validation error as test error.

Answer Point Value: 1.0 points

Answer Key: D

We are given training data $(\mathbf{x}_t, y_t)_{t=1, \dots, N}$ for fitting. We have $n = 10$ features and $N = 1000$ samples. We consider following methods, for fitting these data.

	Parameters	Training error	Validation error
Ridge regression, $\lambda = 1$	10	2.51	3.52
Kernel Ridge regression, $\lambda = 1$	1000	0.53	2.20
Two-layer Neural network with 10 hidden neurons	121	1.56	1.99
Ten-layer Deep Neural network with 10 neurons in each hidden layer	1121	0.22	2.52

Which of these models is preferable and why? Which model would you explore further and how?

- ☐ A. Ridge regression
- ☐ B. Kernel Ridge regression
- ☐ C. Two-layer neural network
- ☐ D. Ten-layer neural network

Answer Point Value: 1.0 points

Answer Key: C