"Machine Learning and Computational Statistics"

4^{th} Homework

Exercise 1 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(\mathbf{x}) + \eta$. Which of the following statements are true?

1. If the joint pdf $p(y, \boldsymbol{x})$ is known, then no noise is involved in the regression task; that is $\eta = 0$.

2. If the joint pdf $p(y, \boldsymbol{x})$ is available, then $g(\cdot)$ can be uniquely estimated by $p(y, \boldsymbol{x})$, through the minimization of the MSE criterion.

3. If a finite data set $Y = \{(y_n, \boldsymbol{x}_n),\ n = 1, \dots, N\}$ is available, then it is possible to compute the exact noise value associated with each data pair in Y, using e.g., the Least Squares criterion.

4. In practice, when a finite data set $Y = \{(y_n, \boldsymbol{x}_n),\ n = 1, \dots, N\}$ is available, then a model of $g(\cdot)$, parameterized by a vector, $\boldsymbol{\theta}$**,** is adopted, and $\boldsymbol{\theta}$ is estimated based on Y, using, e.g., the Least Squares criterion.

5. In the case of finite data sets, when a model of $g(\cdot)$, parameterized by a vector, $\boldsymbol{\theta}$**,** is adopted, then different estimates of $\boldsymbol{\theta}$ are obtained for different data sets, using, e.g., the Least Squares criterion.

Exercise 2 (multiple choices question): Consider the general nonlinear regression task, $\mathrm{y} = g(\mathbf{x}) + \eta$, and assume that the joint pdf, $p(y, \boldsymbol{x})$, is known. Which of the following statements are true?

1. The optimal MSE estimate $\hat{y} := \hat{g}(\boldsymbol{x})$, given an observation $\mathbf{x} = \boldsymbol{x}$, is the solution of the problem: $argmin_{f:R^l \to R} \int_{-\infty}^{+\infty} \left(y - f(\boldsymbol{x})\right)^2 p(y|\boldsymbol{x}) dy$

2. The optimal MSE estimate $\hat{y} := \hat{g}(\boldsymbol{x})$, given an observation $\mathbf{x} = \boldsymbol{x}$, equals to $\int_{-\infty}^{+\infty} \left(y - f(\boldsymbol{x})\right)^2 p(y|\boldsymbol{x}) dy$

3. The estimate $\hat{y} := \hat{g}(\boldsymbol{x}) = \mathrm{E}[y|\boldsymbol{x}]$ is also optimum with respect any other criterion different than the MSE.

4. The optimal MSE estimate $\hat{y} := \hat{g}(\boldsymbol{x})$, given an observation $\mathbf{x} = \boldsymbol{x}$, equals to $\int_{-\infty}^{+\infty} y p(y|\boldsymbol{x}) dy$

Exercise 3 (multiple choices question): Consider the general nonlinear regression task, $\mathrm{y} = g(\mathrm{x}) + \eta$, where both y and x are (scalar) random variables, and assume that the joint pdf

$p(y,x)$ is Gaussian, with mean $\boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} {\sigma_y}^2 & \sigma_{yx} \\ \sigma_{yx} & {\sigma_x}^2 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 1 & 9 \end{bmatrix}$. It turns out that, in the case of normal pdf, the optimum MSE estimate for $g(\cdot)$ is $\hat{g}(x) = E[y|x] = \mu_y + \frac{a\sigma_y}{\sigma_x}(x - \mu_x)$, where $a = \frac{\sigma_{yx}}{\sigma_x \sigma_y}$. Which of the following statements is/are true?

1. If $x = 1$, then $\hat{y} = \frac{38}{9}$.

2. If $x = 2$, then $\hat{y} = 4$.

3. If $x = 3$, then $\hat{y} = \frac{37}{9}$.

4. If $x = 4$, then $\hat{y} = \frac{35}{9}$.

Exercise 4 (multiple choices question): Consider the general nonlinear regression task, $\mathrm{y} = g(\mathbf{x}) + \eta$, where η is zero mean i.i.d. Gaussian noise with variance ${\sigma_\eta}^2$, and let $\hat{y} = \mathrm{E}[\mathrm{y}|\boldsymbol{x}]$ be the related MSE estimate of y, given $\boldsymbol{x}$. Which of the following statements are true?

1. There may exist another estimate that exhibits lower MSE value than that of $\hat{y}$.

2. The MSE value corresponding to $\hat{y}$ equals to ${\sigma_\eta}^2$.

3. There exist other estimates, whose corresponding MSE value may be smaller than ${\sigma_\eta}^2$.

Exercise 5 (multiple choices question): Consider the general nonlinear regression tasks, $\mathrm{y} = g(\mathrm{x}) + \eta_y$ and $\mathrm{z} = g(\mathrm{x}) + \eta_z$, where the respective joint pdfs are Gaussians, i.e., $p(y,x) = N(\boldsymbol{\mu}, \Sigma_1) = N\left(\begin{bmatrix} \mu \\ \mu_x \end{bmatrix}, \begin{bmatrix} {\sigma_y}^2 & \sigma_{yx} \\ \sigma_{yx} & {\sigma_x}^2 \end{bmatrix}\right)$ and $p(z,x) = N(\boldsymbol{\mu}, \Sigma_2) = N\left(\begin{bmatrix} \mu \\ \mu_x \end{bmatrix}, \begin{bmatrix} {\sigma_z}^2 & \sigma_{zx} \\ \sigma_{zx} & {\sigma_x}^2 \end{bmatrix}\right)$, respectively (that is, both pdfs share the same mean vector and they differ in their covariance matrices). Recall that, in the case of normal pdfs, the optimum MSE estimate for $g(\cdot)$ is $\hat{g}(x) = \mathrm{E}[y|x] = \mu_y + \frac{a\sigma_y}{\sigma_x}(x - \mu_x)$, where $a = \frac{\sigma_{yx}}{\sigma_x \sigma_y}$. Let $\hat{y} = \mathrm{E}[\mathrm{y}|x]$ and $\hat{z} = \mathrm{E}[\mathrm{z}|x]$ be the MSE estimates associated with y and z, given x, respectively. Which of the following statements are true?

1. For a given x, it is, in general, $\hat{y} \neq \hat{z}$.

2. Although for a given x, it may be $\hat{y} = \hat{z}$, the MSE values associated with $\hat{y}$ and $\hat{z}$ will differ, if ${\sigma_y}^2 \neq {\sigma_z}^2$ and $\sigma_{yx} \neq \sigma_{zx}$.

3. If $\hat{y}$ is the estimate corresponding to x_1, and $\hat{z}$ the estimate corresponding to $x_2(\neq x_1)$, then $\hat{y} \neq \hat{z}$, when both Σ_1 and Σ_2 are diagonal.

4. For any choice of Σ_1 and Σ_2, for $x = \mu_x$, it is $\mathrm{E}[\mathrm{y}|x] = \mathrm{E}[\mathrm{z}|x]$.

5. For any x, it holds generally that the MSE values associated with $\hat{y}$ and $\hat{z}$ will coincide.

Exercise 6 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(\mathrm{x}) + \eta$, where both y and x are (scalar) random variables and η is zero mean i.i.d. Gaussian noise with variance ${\sigma_\eta}^2 = 5$. In addition, the joint distribution, $p(y, x)$, is normal with mean $\boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} = \begin{bmatrix} 6 \\ 7 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} {\sigma_y}^2 & \sigma_{yx} \\ \sigma_{yx} & {\sigma_x}^2 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 1 & 9 \end{bmatrix}$. Let $\hat{y} = \mathrm{E}[\mathrm{y}|x]$ be the related MSE estimate of y, given x, respectively. Which of the following statements are true?

1. The MSE estimate of y, given x, is a linear function of x.

2. The MSE estimate of y, given x, is equal to 4.

3. The MSE estimate of x, given y, is equal to 1.

4. The variance of y around the optimal MSE estimate, $\hat{y}$, is equal to 5.

Exercise 7 (multiple choices question): Consider the regression task $\mathrm{y} = g(\mathrm{x}) + \eta$, where both y and x are (scalar) random variables and η is zero mean i.i.d. Gaussian noise, whose variance is ${\sigma_\eta}^2$. Fix $x = 1$. Which of the following statements are true?

1. The variance of y around the optimal MSE estimate is less than ${\sigma_\eta}^2$.

2. The MSE estimate of y, given $x = 1$, is equal to $\mathrm{E}[y|1]$.

3. The joint pdf of y and x is a multivariate Gaussian function.

4. The regression task can be expressed as $\mathrm{y} = A + \eta$, where A is a constant.

Exercise 8 (multiple choices question): Consider the regression task $\mathrm{y} = g(\mathrm{x}) + \eta$, where both y and x are (scalar) random variables and for their joint pdf it holds $p(y, x) = p(y) \cdot p(x)$. Then, fixing x to a certain value x, the optimal MSE estimate, $\mathrm{E}[\mathrm{y}|x]$, equals to:

1. $\mathrm{E}[\mathrm{y}]$

2. $x \cdot \mathrm{E}[\mathrm{y}]$

3. $p(x) \cdot \mathrm{E}[\mathrm{y}]$

4. $\frac{1}{p(x)} E[y]$

<u>Exercise 9 (multiple choices question):</u> Consider the regression task $\mathrm{y} = g(\mathrm{x}) + \eta$, where both y and x are (scalar) random variables and η is the noise with mean equal to $a (\neq 0)$ and variance equal to zero (an utopic case – it may model a systematic error involved in the measurements). Recall that the optimal MSE estimate for y, given a specific value x, is equal to $\mathrm{E}[\mathrm{y}|x]$. Which of the following statements are correct?

1. $\mathrm{E}[\mathrm{y}|x] = g(x) + a$

2. $\mathrm{y} + g(\mathrm{x}) = \alpha$

3. $\mathrm{E}[(\mathrm{y} - \mathrm{E}[\mathrm{y}|x])^2] = 0$

4. $\mathrm{E}[(g(x) - \mathrm{E}[\mathrm{y}|x])^2] = 0$

<u>Exercise 10 (multiple choices question):</u> Consider the general nonlinear regression task $\mathrm{y} = g(\boldsymbol{x}) + \eta$, where $\boldsymbol{x}$ is the input, y is the associated output random variable and η is the random noise ($\boldsymbol{x}$ is considered fixed to a specific value). Which of the following statements are true?

1. The optimum estimate $\hat{y}$ associated with $\boldsymbol{x}$, in the MSE sense, equals to $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$.

2. The optimum estimate $\hat{y}$, associated with $\boldsymbol{x}$, in the MSE sense, can always be computed analytically.

3. The computation of the quantity $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$ requires knowledge of the joint pdf $p(y, \boldsymbol{x})$, which, in practice, is always analytically available.

4. Even if the joint pdf $p(y, \boldsymbol{x})$ is available, the computation of $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$ may not be computationally tractable.

<u>Exercise 11 (multiple choices question):</u> Consider the general nonlinear regression task $\mathrm{y} = g(\boldsymbol{x}) + \eta$, where $\boldsymbol{x}$ is the input, y is the associated output random variable and η is the random noise ($\boldsymbol{x}$ is considered fixed to a specific value). Which of the following statements are true?

1. A suboptimal, in the MSE sense, estimate, $\hat{y}$, associated with $\boldsymbol{x}$, may achieve the same MSE value with the optimum one, $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$.

2. In practice, the estimate returned by a suboptimal, in the MSE, estimator is not based on the joint pdf $p(y, \boldsymbol{x})$.

3. In practice, the estimate returned by a suboptimal, in the MSE sense, estimator may be optimal with respect to another optimality criterion.

4. The adoption of a suboptimal estimator is dictated by the fact that a) we do not know the true joint pdf $p(y, \boldsymbol{x})$ and/or b) the involved computations are intractable.

<u>Exercise 12 (multiple choices question):</u> Consider the general nonlinear regression task $\mathrm{y} = g(\boldsymbol{x}) + \eta$, where $\boldsymbol{x}$ is the input, y is the associated output random variable and η is the random noise ($\boldsymbol{x}$ is considered fixed to a specific value). Consider an estimate $f(\boldsymbol{x}; D)$, where D is a specific training data set of size N. The questions are with respect to the bias-variance tradeoff formula. Which of the following statements are true?

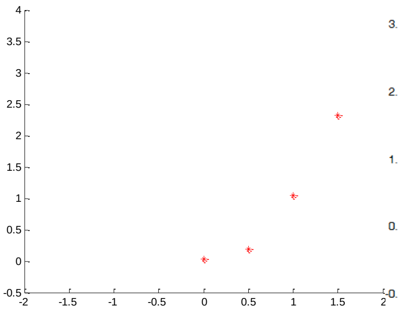
1. The mean in the expression for the mean square deviation of $f(\boldsymbol{x}; D)$ from $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$ is taken over all data sets of the same size, say N.

2. The variance term involved in the expression of the mean square deviation of $f(\boldsymbol{x}; D)$ from $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$, involves the quantity $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$.

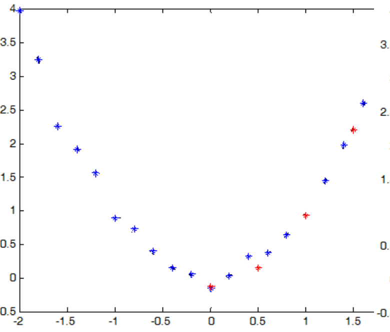
3. The bias term involved in the expression of the mean square deviation of $f(\boldsymbol{x}; D)$ from $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$, quantifies the variance of $f(\boldsymbol{x}; D)$ around $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$.

4. The higher the mean square deviation of $f(\boldsymbol{x}; D)$ from $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$, the more accurate the estimates $f(\boldsymbol{x}; D)$ are.

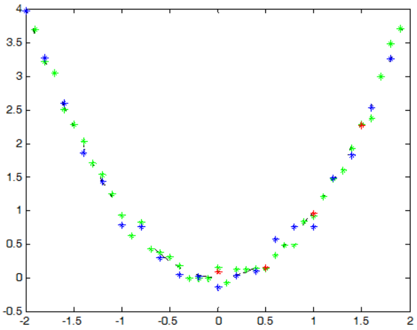
<u>Exercise 13 (multiple choices question):</u> Consider the generalized linear regression task $\mathrm{y} = \boldsymbol{\theta}^T \boldsymbol{\varphi}(x) + \eta$, where x is the (scalar) input, y is the associated output random variable and η is the random noise. The components of the vector $\boldsymbol{\varphi}$ are monomial (power) functions of x. Concerning the associated data set, consider the following three scenarios: (i) the data set D_1, with N_1points, in Fig. (a), (ii) the data set D_2, with $N_2 > N_1$ points, in Fig. (b) ($D_1 \subset D_2$, with the extra points in D_2 being colored blue), and (iii) the data set D_3, with $N_3 > N_2$ points, in Fig. c ($D_2 \subset D_3$, with the extra points in D_3 being colored green). The problem is first to define $\boldsymbol{\varphi}(\cdot)$, for each scenario, and then to estimate the associated $\boldsymbol{\theta}$, based on a set of data points. Match each one of the three scenarios (left column) with the least complex $\boldsymbol{\varphi}(\cdot)$ function (right column) that matches the data.



(a)



(b)



(c)

(a)	1. $\boldsymbol{\varphi}(\boldsymbol{x}) = [1, x]^T$
(b)	2. $\boldsymbol{\varphi}(\boldsymbol{x}) = [1, x, x^2]^T$
(c)	3. $\boldsymbol{\varphi}(\boldsymbol{x}) = [1, x, x^2, x^3]^T$

Exercise 14 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(x) + \eta$, where x is the (scalar) input, y is the associated output random variable and η is the random noise (x is considered fixed to a specific value). Assume that the optimal MSE estimate corresponding to x, $\mathrm{E}[\mathrm{y}|x]$, is a quadratic function of x. Let $f(x; D)$ be another estimate of y. Which of the following statements are true?

1. If $f(x; D)$ is quadratic, then the bias term associated with the mean square deviation of $f(x; D)$ from $\mathrm{E}[\mathrm{y}|x]$, may become zero.

2. If $f(x; D)$ is a tenth degree polynomial, and D comprises 11 points, then the variance term associated with the mean square deviation of $f(x; D)$ from $\mathrm{E}[\mathrm{y}|x]$, is expected to have a very small positive value.

3. If $f(x; D)$ is a linear function, then the bias term associated with the mean square deviation of $f(x; D)$ from $\mathrm{E}[\mathrm{y}|x]$, will definitely be zero.

4. If $f(x; D)$ is a tenth degree polynomial, and D comprises 11 points, then the bias term associated with the mean square deviation of $f(x; D)$ from $E[\mathrm{y}|x]$, is expected to be very close to zero.

Exercise 15 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(x) + \eta$, where x is the (scalar) input that takes values in an interval in the real axis, y is the associated output random variable and η is the (zero mean) random noise (x is considered fixed to a certain value). Assume that the optimal MSE estimate corresponding to x, $\mathrm{E}[\mathrm{y}|x]$, is a 3rd order/degree polynomial function (of course, in practice, this information is rarely known).

Let $f(\cdot;D)$ be another estimate of y, associated with the data set D, of size N. Which of the following scenarios for $f(\cdot;D)$ is expected to approximate better $\mathrm{E}[\mathrm{y}|x]$?

1. $f(\cdot;D)$ is linear and $N = 100000$.

2. $f(\cdot;D)$ is quadratic and $N = 1000$.

3. $f(\cdot;D)$ is a tenth degree polynomial and $N = 10$.

4. $f(\cdot;D)$ is a fourth degree polynomial and $N = 300$.

Exercise 16 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(x) + \mathrm{\eta}$, where x is the (scalar) input, y is the associated output random variable and η is the (zero mean) random noise (x is considered fixed to a specific value). Assume that the optimal MSE estimate corresponding to x is $\mathrm{E}[\mathrm{y}|x] = 2 \cdot x + 3$. Let $f(x;D)$ be another estimate of y, associated with the data set D, of size N. Which of the following statements are true?

1. If $f(x;D) = 3$, the variance term involved in the expression of the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$, is zero.

2. If $f(x;D) = 2 \cdot x + 3$, the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$, is zero.

3. There exists an $f(x;D)$, for which the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$, is exactly equal to zero, for finite N.

4. If $f(x;D) = 3$, the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$, decreases, as N increases.

Exercise 17 (multiple choices question): Consider the general nonlinear regression task $\mathrm{y} = g(x) + \mathrm{\eta}$, where x is the (scalar) input that takes values in an interval in the real axis, y is the associated output random variable and η is the (zero mean) random noise (x is considered fixed to a specific value). Assume that the noise variance is approximately equal to zero (${\sigma_\eta}^2 \approx 0$) and the optimal MSE estimate corresponding to x is $\mathrm{E}[\mathrm{y}|x] = 4 \cdot x - 3$. Let $f(x;D)$ be another estimate of y, associated with the data set D, of size $N = 3$. Which of the following statements are true?

1. If $f(x;D) = \theta_1 \cdot x + \theta_0$, where $\theta_1$and θ_0 are estimated via the LS criterion based on D, the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|\boldsymbol{x}]$, will be approximately equal to zero.

2. If $f(x;D) = 3$, the bias term involved in the expression of the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$, is zero.

3. If $f(\cdot;D) = \theta_2 \cdot x^2 + \theta_1 \cdot x + \theta_0$, where $\theta_2, \theta_1$and θ_0 are estimated via the LS criterion based on D, the mean square deviation of $f(x;D)$ from $\mathrm{E}[\mathrm{y}|x]$,is approximately equal to zero, since the estimate of θ_2 will be very close to zero.

4. If $f(\cdot;D) = \theta_0$, where θ_0 is estimated via the optimization of the LS criterion based on D, the mean square deviation of $f(x;D)$ from $E[\mathrm{y}|x]$, approximates the zero value.

Exercise 18 (regularization - python code):

Consider the data set given in the attached file (the code for reading from python is also given). Specifically, it consists of 10 data pairs of the form (y_i, x_i), $i = 1, \ldots, 10$. All y_i's are accumulated in the vector $\boldsymbol{y}$ while all x_i's are accumulated in the vector $\boldsymbol{x}$.

The aim is to unravel the relation between x_i's and y_i's.

(a) Plot the data.

(b) Fit a 8th degree polynomial on the data using the LS estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial.

(c) Fit a 8th degree polynomial on the data using the ridge regression estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial. Experiment with various values of λ.

(d) Fit a 8th degree polynomial on the data using the lasso estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial. Experiment with various values of λ.

(e) Discuss briefly on the results produced by lasso and compare them with those produced by the LS and ridge regression cases.

Exercise 19 (python code + text):

Consider the set-up of exercise 15 from Homework 3. Consider also the ridge regression estimators resulting from eq. (A) in the exercise 1 above, for $\lambda = 0, 0.1, 0.2, \ldots, 10000$. For each one of these values of λ, apply the steps (a), (b), (c1) of exercise 15 above, in order to compute the MSE. Plot MSE versus λ and

(i) determine the range of values of λ where the MSE is smaller than that of the unbiased LS estimator,

(ii) Comment on the results.

Exercise 20:

Consider the **regression problem** $\mathrm{y} = g(\mathbf{x}) + \eta$

It is known that $E[\mathrm{y}|\boldsymbol{x}]$ is the minimum MSE estimate of y given $\boldsymbol{x}$. Consider the estimator $f(\boldsymbol{x}; D)$. In the light of the discussion that took place during the 4th lecture, answer the following questions:

(a) Under what conditions (theoretically) the quantity $E_D[(f(\boldsymbol{x}; D) - E[\mathrm{y}|\boldsymbol{x}])^2]$ becomes zero?

(b) Why this cannot be achieved in practice?

Exercise 21:

Consider a regression task $y = g(x) + \eta$, where y and x are modeled by the random variables y and x. The joint pdf of y and x is:

$p(x, y) = \frac{4}{3}$, for $x \in (0,1), y \in (x^3, 1)$.

Determine the optimum MSE estimate $E[y|x]$, for a given x, by performing the following steps:

(a) Verify that $p(x, y)$ is a pdf (prove that it integrates to 1).

(b) Compute the marginal pdf of x, $p_{\mathrm{x}}(x)$.

(c) Compute the conditional pdf $p(y|x)$.

(d) Compute and plot $E[y|x]$.

Hint: It is $\int_a^b x^n dx = [\frac{1}{n+1} x^{n+1}]_a^b = \frac{1}{n+1} b^{n+1} - \frac{1}{n+1} a^{n+1}$

Exercise 22 (python code + text):

Consider the **regression problem** (1-dep., 1-indep. variables)

$$\mathrm{y} = g(\mathrm{x}) + \eta$$

where y and x are jointly distributed according to the normal distribution $p(y, x) = N(\boldsymbol{\mu}, \Sigma)$

with $\boldsymbol{\mu} = \begin{bmatrix}\mu_y \\ \mu_x\end{bmatrix} = \begin{bmatrix}1 \\ 1\end{bmatrix}$ and $\Sigma = \begin{bmatrix}{\sigma_y}^2 & \sigma_{yx} \\ \sigma_{yx} & {\sigma_x}^2\end{bmatrix} = \begin{bmatrix}4 & 3 \\ 3 & 5\end{bmatrix}$

(a) Determine $E[\mathrm{y}|\boldsymbol{x}]$ and plot the corresponding curve (recall the relevant theory concerning the normal distribution case).

(b) Generate 100 data sets D_i, $i = 1, \dots 100$, each one consisting of $N = 50$ randomly selected pairs (y_n, x_n), n=1,…,N, from $p(y, x)$.

(c) Adopt a linear estimator $f(x; D)$ and determine its instances $f(x; D_1), \dots, f(x; D_{100})$, utilizing the LS criterion.

(d) Plot in a single figure **(i)** the lines corresponding to the above 100 estimates (blue color) and **(ii)** the line corresponding to the optimal MSE estimate (green color).

(e) Repeat steps (b)-(d) where now each data set consists of $N = 5000$ points.

(f) Discuss the results (in your discussion, take into account the decomposition of the MSE to a variance and a bias term).

Exercise 23 (python code + text)**:**

Consider the set-up of exercise 21 and recall the $E[\mathrm{y}|\boldsymbol{x}]$ determined there.

(a) Generate a single data set D of 100 pairs (y_n, x_n), $n = 1, \dots, 100$ from $p(y, x)$.

(b) Determine the linear estimate $f(x; D)$ that minimizes the MSE criterion, based on D.

(c) Generate randomly a set D' of additional 50 points (y'_n, x'_n), $n = 1, \dots, 50$. For each x'_n determine the estimate $y'_n = f(x_n; D')$ (50 numbers (estimates) should be finally computed).

(d) Again, for the 50 x'_n 's determine the associated estimates $\hat{y} = E[y|\boldsymbol{x}]$.

(e) Based on the previous derived estimates for the 50 points from both $f(x_n; D)$ and $E[y|\boldsymbol{x}]$, propose and use a (practical) way for quantifying the performance of the two estimators $f(x_n; D')$ and $E[y|\boldsymbol{x}]$.

Exercise 24 (python code + text)**:** Consider the setup of exercise 21. Generate a set D of $N = 100$ data pairs $\boldsymbol{z}_n = (y_n, x_n)$.

(a) For each x_n compute the optimal MSE estimate (use the results of exercise 3).

(b) Compute $\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} x_n \\ \frac{1}{N}\sum_{n=1}^{N} y_n \end{bmatrix}$ and $\Sigma = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\mu} - \boldsymbol{z}_n)(\boldsymbol{\mu} - \boldsymbol{z}_n)^T$.

(c) Pretend that you do not know the true distribution that generates the data and you (erroneously) assume that the joint pdf of x and y is a normal one with mean and covariance matrix those computed in (b). Derive the optimum MSE estimate for this case and compute the MSE estimate for each one of the 100 x_n 's.

Discuss the results obtained from (a) and (c).