

**Exercise 2: Statistics + Parameter Estimation**  
(to be returned on November 7th, 9:00)

Prof. Dr. Moritz Diehl, Katrin Baumgärtner, Jakob Harzer, Yizhen Wang, Rashmi Dabir

---

In this exercise you get to know some matrix properties. In addition, you investigate some important facts from statistics in numerical experiments. Pen-and-paper exercises can be uploaded on the Ilias course page as a digitally created PDF or handed in during the lecture.

**Exercise Tasks**

1. PAPER: The covariance matrix of a vector-valued random variable  $X \in \mathbb{R}^n$  with mean  $\mathbb{E}\{X\} = \mu_X$  is defined by

$$\text{cov}(X) := \mathbb{E}\left\{(X - \mu_X)(X - \mu_X)^\top\right\}.$$

Prove that the covariance matrix of a vector-valued variable  $Y = AX + b$  with constant  $A \in \mathbb{R}^{m \times n}$  and  $b \in \mathbb{R}^m$  is given by

$$\text{cov}(Y) = A \text{cov}(X) A^\top.$$

(2 points)

2. PAPER: Let  $X \in \mathbb{R}^n$  be a vector-valued random variable with mean  $\mu \in \mathbb{R}^n$ . Show that the covariance matrix  $\text{cov}(X)$  can also be calculated by

$$\text{cov}(X) = \mathbb{E}\{XX^\top\} - \mu\mu^\top$$

(2 points)

3. PAPER: Suppose we are measuring a constant  $x_0 \in \mathbb{R}$  perturbed by random independent noise  $\epsilon$  with mean  $\mu_\epsilon = 0$  and variance  $\sigma_\epsilon^2 > 0$ , i.e. we have

$$x = x_0 + \epsilon.$$

- (a) State the mean  $\mu_x$  and the variance  $\sigma_x^2$  of the random variable  $x$ . (1 point)
- (b) Let  $x(n) = (x_1, \dots, x_n)$  denote a sample of  $n$  observations of  $x$ . The sample mean is given by  $\bar{x}(n) = \frac{1}{n} \sum_{i=1}^n x_i$  and it is an unbiased estimator of the mean  $\mu_x$ . What is the variance of  $\bar{x}(n)$ ? (1 point)
- (c) Prove that the Least Squares (LS) estimate for  $x_0$  is the sample mean  $\bar{x}(n)$ . State the minimization problem explicitly. Is it convex? (2 bonus points)

4. Consider the following experimental setup, where we measure the temperature-dependent expansion of a steel bar. Here  $L_0$  [cm] is the length of the bar at the beginning of the experiment and  $L(T)$  [cm] represents the length of the bar at temperature  $T$  [K]. The following relationship holds, between the length of the bar at temperature  $T_0$  [K]:  $L_0 = L(T_0)$ . We define  $\Delta T := T - T_0$  as the independent variable. Furthermore, we define  $A := \alpha \cdot L_0$  [cm/K], where  $\alpha$  [1/K] is the specific expansion coefficient. Then the model is given by

$$m(\Delta T(k); A, L_0) = A \cdot \Delta T(k) + L_0. \quad (1)$$

Below, you find the datapoints. Using the data, you will compute estimates for the parameters  $A$  and  $L_0$ .

$k$	1	2	3	4
$\Delta T(k)$ [K]	5	15	35	60
$L(k)$ [cm]	6.55	9.63	17.24	29.64

- (a) CODE: Plot the  $\Delta T(k), L(k)$  relation using 'x' markers. (0.5 points)
- (b) PAPER: Using the model from above, calculate the experimental values for the parameters  $A$  and  $L_0$  by minimizing the sum of squared distances, i.e.

$$A^*, L_0^* = \arg \min_{A, L_0} \sum_{k=1}^4 d_k(A, L_0)^2, \quad (2)$$

where the distance  $d_k$  is given by

$$d_k(A, L_0) = L(k) - m(\Delta T(k); A, L_0).$$

CODE: Plot the fit  $m(\Delta T; A^*, L_0^*) = A^* \Delta T + L_0^*$  over the range  $[0, 100]$  in the same figure as before.

*Hint: Compute the solution by setting the gradient of the objective function with respect to the parameters  $(A, L_0)$  to zero, i.e.  $\nabla_{(A, L_0)} \sum_k d_k^2 = 0$ . This will give you a  $2 \times 2$  linear system. Check if the objective function is convex!* (2 points)

- (c) CODE: Now, use a third order polynomial and fit it to the data using `np.polyfit`. Again minimize the sum of squared distances to find optimal values for the coefficients of your model equation. Plot the fit in the same figure as before. (0.5 point)
- (d) CODE: You take another measurement: at  $\Delta T = 70$ K you measure a length of  $L = 32.89$  cm. You can use this additional datapoint to validate your fit. Therefore plot it in the existing plot. PAPER: Which fit looks more reasonable to you?

*Hint: The phenomenon of fitting a model to a data set which then does not pass validation is called 'overfitting'.* (1 point)

*This sheet gives in total 10 points and 2 bonus points.*