Bonus Assignment – Least Squares for Real-World Problems
Due date: Wednesday, 22 November 2023, 11:59 PM

Exercise 1: Inconsistent systems of equations [10 points]

Consider the following inconsistent systems of equations:

$$\text{(a) } A_1\mathbf{x} = \mathbf{b_1},\ \text{with } A_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \mathbf{b_1} = \begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix}$$

$$\text{(b) } A_2\mathbf{x} = \mathbf{b_2},\ \text{with } A_2 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \mathbf{b_2} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Find the least squares solution $\mathbf{x}^*$ and compute the Euclidean norm of the residual, SE and RMSE.

Exercise 2: Polynomials models for least squares [20 points]

In this exercise, we consider two small datasets about the crude oil (`crudeOil.txt`) and kerosene (`kerosene.txt`) production by year in Europe in the period 1980-2012. By solving the following tasks, we will try to fit the data with different polynomial models and determine the best one.

(a) Write a Matlab function `leastSquares.m` which takes, as inputs, matrix A and vector $\mathbf{b}$ of a generic system $A\mathbf{x} = \mathbf{b}$ and returns, as outputs, the least squares solution $\mathbf{x}^*$, the Euclidean norm of the residual, the SE and the RMSE. In order to test your function, write a script `ex2a.m`, in which you use `leastSquares()` to find the solutions of the two inconsistent systems of Exercise 1 and compare the results obtained with the ones you computed by hand.

(b) Consider the linear model $y_i = \alpha_1 + \alpha_2 x_i$ and apply it to the crude oil and kerosene production data in the period 1980-2011. Write a script `linearModel.m` in which you use `leastSquares()` to compute the least squares solution $\mathbf{x}^*$ and the metrics of the residual. For each dataset, create a figure in which you plot the original data points and the linear model.

(c) Consider the quadratic model $y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2$ and apply it to the crude oil and kerosene production data in the period 1980-2011. Write a script `quadraticModel.m` in which you use `leastSquares()` to compute the least squares solution $\mathbf{x}^*$ and the metrics of the residual. For each dataset, create a figure in which you plot the original data points and the quadratic model.

(d) Consider the cubic model $y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + \alpha_4 x_i^3$ and apply it to the crude oil and kerosene production data in the period 1980-2011. Write a script `cubicModel.m` in which you use `leastSquares()` to compute the least squares solution $\mathbf{x}^*$ and the metrics of the residual. For each dataset, create a figure in which you plot the original data points and the cubic model.

(e) Compare the linear, quadratic and cubic models on the basis of the quality metrics computed above, by creating a table containing the results for the two models. Which one of the three models would you pick for the crude oil data? And for the kerosene? Provide an estimate of the crude oil and kerosene production in 2012 by using the three models and compare the values obtained with the real values reported in the data source. Comment on your results.

Exercise 3: Analysis of periodic data [20 points]

The file `temperature.txt` contains the area mean-temperatures of Switzerland between January 1864 and March 2021 included. Temperature data exhibit a periodic behaviour and we will try to capture it by using periodic models. You will need the function `leastSquares()` implemented in Exercise 2.

(a) Consider the periodic model $y_i = \alpha_1 + \alpha_2 \cos(2\pi x_i) + \alpha_3 \sin(2\pi x_i)$ and apply it to the temperature data: (I) between January 1960 and January 1963; (II) between January 1960 and January 1970. Write a script `periodicA.m` in which you compute the least squares solutions and the metrics of the residual, and plot the outputs of the model against the original data in both cases.

(b) Repeat the same analysis and plots of the previous point for both time series, by using the periodic model $y_i = \alpha_1 + \alpha_2 \cos(2\pi x_i) + \alpha_3 \sin(2\pi x_i) + \alpha_4 \cos(4\pi x_i)$ in the script `periodicB.m`.

(c) Compare the models of point (a) and (b). Was it beneficial to include more data? Which model would you prefer? Are you satisfied with the results obtained? If necessary, what would you suggest to improve your models? Motivate your answers.

Exercise 4: Data linearization and Levenberg-Marquardt method for the exponential model [20 points]

The file `nuclear.txt` contains the data on the nuclear electric power consumption by year in China in the period 1999-2006. We consider the power law model, expressed as:

$$y_i = \alpha_1 x_i^{\alpha_2}.$$

(a) Find the least squares best fit by using data linearization and compute the RMSE both of the log-linearized model and of the original exponential model. Include in your report all the computations and the necessary steps, as explained in the slides of the tutorial.

(b) Write a function `levenbergMarquardt()` in which you implement the Levenberg-Marquardt algorithm for solving nonlinear least squares problems. Following again the slides of the tutorial, show how you can formulate the problem in order to solve it with Levenberg-Marquardt method and compute analytically all the necessary quantities. Finally, write a script `ex4b.m` in which you use the function `levenbergMarquardt()` to fit the data points and compute the RMSE.

(c) Compare the results obtained in points (a) and (b), by extending the script `ex4b.m` to produce a plot of the original data points together with the two models. Which model would you choose?

Exercise 5: Tikhonov regularization [15 points]

Let us consider an ill-posed problem $A\mathbf{x} = \mathbf{b}$, with $A \in \mathbb{R}^{n\times m}$ and $\mathbf{b} \in \mathbb{R}^m$. In order to solve it, we introduce a parameter $\alpha \in \mathbb{R}$ and we use linear least squares with Tikhonov regularization, given by:

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2 . \tag{1}$$

(a) Derive analytically the optimal solution $\mathbf{x}^*$ of Eq. (1). What is the purpose of the parameter α and how should we proceed to choose its value?

(b) We now consider the Hilbert matrix $H \in \mathbb{R}^{n\times n}$, with entries defined as follows:

$$H_{ij} = \frac{1}{i+j-1} ,$$

for every $i, j = 1, \ldots, n$. Write a Matlab script `illposedHilbert.m` in which you generate H for $n = 50, 100, 200, 300, 400, 500, 1000$ and solve the problem $H\mathbf{x} = \mathbf{b}$, for $b = H\mathbf{x}_{\text{exact}}$ and $\mathbf{x}_{\text{exact}}$ generated through the function `rand(n,1)`. To make your results reproducible, reinitialize the random number generator to its startup configuration by adding `rng('default')` at the beginning of your script. Produce also two figures in which you plot: (I) the condition number of H (use `cond()` in Matlab) against n; (II) the norm of the error $\|\mathbf{x}_{\text{exact}} - \mathbf{x}\|_2$ against n.

(c) We now focus our attention on the case $n = 100$. Write a Matlab script `regularizedHilbert.m` in which you estimate the regularized solution $\mathbf{x}_{\text{reg}}$ according to Eq. (1) by using at least 10 different values of the parameter α (explain your choice of the values). To visualize the results, produce two figures in which you plot: (I) the norm of the error $\|\mathbf{x}_{\text{exact}} - \mathbf{x}_{\text{reg}}\|_2$ against the values of α; (II) $\|H\mathbf{x} - \mathbf{b}\|_2$ against $\|\mathbf{x}\|_2$ for the different values of α. Comment your results.

Quality of the code and of the report [15 points]

The highest possible score of each project is 85 points and up to 15 additional points can be awarded based on the quality of your report and code (maximum possible grade: 100 points). Your report should be a coherent document, structured according to the template provided on iCorsi. If there are theoretical questions, provide a complete and detailed answer. All figures must have a caption and must be of sufficient quality (include them either as .eps or .pdf). If you made a particular choice in your implementation that might be out of the ordinary, clearly explain it in the report. The code you submit must be self-contained and executable, and must include the set-up for all the results that you obtained and listed in your report. It has to be readable and, if particularly complicated, well-commented.

Additional notes and submission details

Summarize your results and experiments for all exercises by writing an extended LaTeX report, by using the template provided on iCorsi (`https://www.icorsi.ch/course/view.php?id=16926`). Submit your gzipped archive file (tar, zip, etc.) – named `project_number_lastname_firstname.zip/tgz` – **on iCorsi** (see *[NC 2023] Project 3 – Submission Spectral Graph Clustering*) **before the deadline**. Submission by email or through any other channel will not be considered. Late submissions will not be graded and will result in a score of 0 points for that project. You are allowed to discuss all questions with anyone you like, but: (i) your submission must list anyone you discussed problems with and (ii) you must write up your submission independently. Please remember that plagiarism will result in a harsh penalization for all involved parties.

In-class assistance

If you experience difficulties in solving the problems above, please join us in class either on Tuesdays 15:30-17:00 or on Wednesdays 13:30-15:00 in room C1.03.