Lab Session 10

MA-581 : Numerical Computations Lab R. Alam Date: October 28, 2025

The purpose of this lab tutorial is to solve the Least-Squares Problem (in short, LSP) $Ax \simeq b$. Here $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$, and usually m is much bigger than n.

Origin: Suppose that we have a data set (t_i, b_i), for $i = 1 : m$, that have been obtained from some experiment. These data are governed by some unknown laws. So, the task is to come up with a model that best fits these data. A model is generated by a few functions, called model functions, $\phi_1, \ldots, \phi_n$. Therefore once a model is chosen, the task is to find a function p from the span of the model functions that best fits the data.

Suppose that the model functions $\phi_1, \ldots, \phi_n$ are given. For $p \in \text{span}(\phi_1, \ldots, \phi_n)$, we have $p = x_1\phi_1 + \cdots + x_n\phi_n$ for some $x_j \in \mathbb{C}$. Now, forcing p to pass through the data (t_i, b_i) for $i = 1 : m$, we have $p(t_i) = b_i + r_i$, where r_i is the error. We want to choose that p for which the sum of the squares of the errors r_i is the smallest, that is, $\sum_{i=1}^m |r_i|^2$ is minimized.

Now $p(t_i) = b_i + r_i$ gives $x_1\phi_1(t_i) + \cdots + x_n\phi_n(t_i) = b_i + r_i$. Thus in matrix notation,

$$\begin{bmatrix} \phi_1(t_1) & \cdots & \phi_n(t_1) \\ \phi_1(t_2) & \cdots & \phi_n(t_2) \\ \vdots & \cdots & \vdots \\ \phi_1(t_m) & \cdots & \phi_n(t_m) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}.$$

This is of the form $Ax = b + r$ and we have to choose $x \in \mathbb{C}^n$ for which $\|r\|_2 = \|Ax - b\|_2$ is minimized. We write this as LSP $Ax = b$.

1. Over 400 years ago Galileo, attempting to find a mathematical description of falling bodies, studied the paths of projectiles. One experiment consisted of rolling a ball down a grooved ramp inclined at a fixed angle to the horizontal, starting the ball at a fixed height h above a table of height 0.778 meters. When the ball left the end of the ramp, it rolled for a short distance along the table and then descended to the floor. Galileo altered the release height h and measured the horizontal distance d that the ball traveled before landing on the floor. The table below shows data from Galileos notes (with measurements converted from puntos to meters)

   Data from Galileos inclined plane experiment.

h	0.282	0.564	0.752	0.948
d	0.752	1.102	1.248	1.410

   Determine a a least squares linear fit $y = c_1 + c_2x$ with the following MATLAB code:

```
%%% Set up data
h = [0.282; 0.564; 0.752; 0.940]; d = [0.752; 1.102; 1.248; 1.410];

%%% Form the 4x2 matrix A and solve for the coefficient vector c.
A = [ones(size(h)), h]; c = A\d;   % solution of the LSP Ac =d
cc = flipud(c);    % order the coefficients in descending
                   % order for polyval
%%% Plot the data points
plot(h,d,'b '), title('Least Squares Linear Fit'), hold
xlabel('release height'), ylabel('horizontal distance')

%%% Plot the line of best fit
hmin = min(h); hmax = max(h); h1 = [hmin:(hmax-hmin)/100:hmax];
plot(h1, polyval(cc, h1), 'r'), axis tight
```

This code produces the line y = 0.4982 + 0.9926x. This line is plotted along with the data points. A measure of the amount by which the line fails to hit the data points is the residual norm $r := \|Ac - d\|_2$. Determine r.

2. Consider the following least squares method for ranking sports teams. Suppose we have four college football teams, called simply T1, T2, T3, and T4. These four teams play each other with the following outcomes:

   - T1 beats T2 by 4 points: 21 to 17.
   - T3 beats T1 by 9 points: 27 to 18.
   - T1 beats T4 by 6 points: 16 to 10.
   - T3 beats T4 by 3 points: 10 to 7.
   - T2 beats T4 by 7 points: 17 to 10.

   To determine ranking points r_1, r_2, r_3, r_4 for each team, we do a least squares fit to the overdetermined linear system:

   $$r_1 - r_2 = 4, r_3 - r_1 = 9, r_1 - r_4 = 6, r_3 - r_4 = 3, r_2 - r_4 = 7.$$

   This system does not have a unique least squares solution, however, since if $[r_1, r_2, r_3, r_4]^\top$ is one solution and we add to it any constant vector then we obtain another vector for which the residual is exactly the same.

   To make the solution unique, we can fix the total number of ranking points, say, at 20. To do this, we add the equation $r_1 + r_2 + r_3 + r_4 = 20$ to those listed above.

   Note that this equation will be satisfied exactly since it will not affect how well the other equalities can be approximated. Determine the values r_1, r_2, r_3, r_4 that most closely satisfy these equations, and based on your results, rank the four teams.

3. Find the polynomial of degree 10 that best fits the function $f(t) = 1/(1 + 25t^2)$ at $30, 50, 100$ equally-spaced points t between -1 and 1. Set up the matrix A and right-hand side vector b, and determine the polynomial coefficients in two different ways:

   (a) By using the MATLAB command `x = A \ b` (which uses a QR decomposition).

   (b) By solving the normal equations $A^\top Ax = A^\top b$. This can be done in MATLAB by typing

   ```
   x = (A' * A) \ (A' * b).
   ```

   Plot the data points, the three polynomials and the function $f(t)$ in a single plot. Compute the residual norm in each case and comment on the results.

   [Note: You can compute the condition number of A or of $A^\top A$ using the MATLAB function `cond`.]

   Now repeat the above experiment for polynomials of degrees $29, 49$ and 99, respectively. One can show that these are unique polynomials which passes through the data points. Plot these polynomials, the data points and the function $f(t)$ in a single plot for $30, 50, 100$ data points. Do the polynomials of higher degree provide a better approximation of $f(t)$ than the polynomials of degree 10 obtained above? To check this, plot all the polynomials and the function $f(t)$ in a single plot for $30, 50$, and 100 data points.

4. Solve the following least-squires problems (linear regression) using the method of normal equation. For each problem, plot the data and their best linear fit in a single plot.

Exercises

5.5.1. Find the straight line $y = \alpha + \beta t$ that best fits the following data in the least squares sense: (*a*)

t_i	-2	0	1	3
y_i	0	1	2	5

(*b*)

t_i	1	2	3	4	5
y_i	1	0	-2	-3	-3

(*c*)

t_i	-2	-1	0	1	2
y_i	-5	-3	-2	0	3

5.5.2. The proprietor of an internet travel company compiled the following data relating the annual profit of the firm to its annual advertising expenditure (both measured in thousands of dollars):

Annual advertising expenditure	12	14	17	21	26	30
Annual profit	60	70	90	100	100	120

(*a*) Determine the equation of the least squares line. (*b*) Plot the data and the least squares line. (*c*) Estimate the profit when the annual advertising budget is \$50,000. (*d*) What about a \$100,000 budget?

5.5.3. The median price (in thousands of dollars) of existing homes in a certain metropolitan area from 1989 to 1999 was:

year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
price	86.4	89.8	92.8	96.0	99.6	103.1	106.3	109.5	113.3	120.0	129.5

(*a*) Find an equation of the least squares line for these data. (*b*) Estimate the median price of a house in the year 2005, and the year 2010, assuming that the trend continues.

5.5.4. A 20-pound turkey that is at the room temperature of 72° is placed in the oven at 1:00 pm. The temperature of the turkey is observed in 20 minute intervals to be $79^\circ, 88^\circ$, and 96°. A turkey is cooked when its temperature reaches 165°. How much longer do you need to wait until the turkey is done?

♡ 5.5.5. The amount of waste (in millions of tons a day) generated in a certain city from 1960 to 1995 was

year	1960	1965	1970	1975	1980	1985	1990	1995
amount	86	99.8	115.8	125	132.6	143.1	156.3	169.5

(*a*) Find the equation for the least squares line that best fits these data.
(*b*) Use the result to estimate the amount of waste in the year 2000, and in the year 2005.
(*c*) Redo your calculations using an exponential growth model $y = ce^{\alpha t}$.
(*d*) Which model do you think most accurately reflects the data? Why?

5.5.6. The amount of radium-224 in a sample was measured at the indicated times.

time in days	0	1	2	3	4	5	6	7
mg	100	82.7	68.3	56.5	46.7	38.6	31.9	26.4

(a) Estimate how much radium will be left after 10 days.
(b) If the sample is considered to be safe when the amount of radium is less than .01 mg, estimate how long the sample needs to be stored before it can be safely disposed of.

5.5.7. The following table gives the population of the United States for the years 1900-2000.

year	1900	1920	1940	1960	1980	2000
population – in millions	76	106	132	181	227	282

(a) Use an exponential growth model of the form $y = ce^{at}$ to predict the population in 2020, 2050, and 3000. (b) The actual population for the year 2020 has recently been estimated to be 334 million. How does this affect your predictions for 2050 and 3000?