Chapter 5

Least-squares Regression

In the chapters on interpolation we discussed the approximation of a given function by another function that exactly reproduced the given data values. We considered both univariate functions (i.e., functions which depend on one variable) and bivariate functions (i.e. functions which depend on two variables). The function that interpolates given data has been called "interpolant". By an "approximant" (or "approximating function") we mean a function which *approximates* the given data values as well as possible (in some sense we need to agree upon later) but does not necessarily *reproduce* the given data values exactly. That is, the graph of the approximant will not in general go through the data points, but will be close to them — this is called *regression*.

A justification for approximation rather than interpolation is the case of experimental or statistical data. The data from experiments are normally subject to errors. The interpolant would exactly reproduce the errors. The approximant, however, allows to adjust for these errors such that a smooth function results. In the presence of noise it would even be foolish and, indeed, inherently dangerous, to attempt to determine an interpolant, because it is very likely that the interpolant oscillates violently about the curve or surface which represents the true function. Another justification is that there may be so many data points that efficiency considerations force us to approximate from a space spanned by fewer basis functions that data points.

In regression, as for interpolation, we consider the case where a function has to be recovered from partial information, e.g. when we only know (possibly noisy) values of the function at a set of points.

5.1 Least-squares basis functions

The most commonly used classes of approximating functions are functions that can be written as linear combinations of basis functions $\varphi_i, i = 1, \ldots, M$, i.e., approximants of type:

$$\phi(\mathbf{x}) = \sum_{i=1}^{M} c_i \, \varphi_i(\mathbf{x}) = \mathbf{c}^T \boldsymbol{\varphi}(\mathbf{x}), \tag{5.1}$$

with

$$\mathbf{c} := \begin{pmatrix} c_1 & c_2 & \ldots & c_M \end{pmatrix}^T, \tag{5.2}$$

and

$$\boldsymbol{\varphi}(\mathbf{x}) := \begin{pmatrix} \varphi_1(\mathbf{x}) & \varphi_2(\mathbf{x}) & \ldots & \varphi_M(\mathbf{x}) \end{pmatrix}^T,$$

exactly the same as for interpolation. In the following we will restrict to this type of approximants. Moreover, we will assume that the function we want to approximate is at least continuous.

Where least-squares differs from interpolation is that the number of basis functions M is in general *less* than the number of data points N,

$$M \leq N,$$

which means we will have more constraints than free variables, and therefore we will not be able to satisfy all constraints exactly.

Commonly used classes of (univariate) basis functions are:

- algebraic polynomials:
$$\varphi_1 = 1, \; \varphi_2 = x, \; \varphi_3 = x^2, \ldots$$
- trigonometric polynomials:
$$\varphi_1 = 1, \; \varphi_2 = \cos x, \; \varphi_3 = \sin x, \; \varphi_4 = \cos 2x, \; \varphi_5 = \sin 2x, \ldots$$
- exponential functions:
$$\varphi_1 = 1, \; \varphi_2 = \mathrm{e}^{\alpha_1 x}, \; \varphi_3 = \mathrm{e}^{\alpha_2 x}, \ldots$$
- rational functions:
$$\varphi_1 = 1, \varphi_2 = \frac{1}{(x-\alpha_1)^{p_1}}, \varphi_3 = \frac{1}{(x-\alpha_2)^{p_2}}, \ldots, \quad p_i \in \mathbb{N}$$

.

For bivariate regression, radial functions are also popular; they have been introduced in the previous chapter.

5.2 Least-squares approximation - Example

In practical applications we are interested in getting "good" approximations, i.e. the approximant should not deviate "much" from the given data. However, what is the precise meaning of "good" and "much"? In other words, how do we measure the quality of the approximation or, equivalently, how do we measure the error in the approximation?

In order to answer this question let us first discuss a simple example. Suppose we are given the data in table 5.1. We have 3 data points. Let us look for a straight line, which best fits in some sense the given data. The equation of the straight line is

i	1	2	3
x_i	0	1	2
f_i	4.5	3.0	2.0

Table 5.1: Example data set.

$$\phi(x) = a_0 + a_1\, x,$$

with so far unknown coefficients a_0 and a_1. Intuitively, we would like the straight line to be as close as possible to the function $f(x)$ that generates the data. A measure of the 'closeness' between $\phi(x)$ and $f(x)$ could be based on the difference of the function values of f and ϕ at the given data points, i.e., on the quantities

$$r_i := f(x_i) - \phi(x_i) = f_i - (a_0 + a_1\, x_i), \quad i = 1, 2, 3.$$

The residuals are shown graphically in Figure 5.1 as the vertical red bars capturing the distance between samples of f at the nodes, and the approximation ϕ. The r_i's are called the *residuals*. The least-squares method finds among all possible coefficients a_0 and a_1 the pair that minimizes the *square sum of the residuals*,

$$\sum_{i=1}^{3} r_i^2,$$

i.e., that makes $\sum_{i=1}^{3} r_i^2$ as small as possible. This minimization principle is sometimes called the *(discrete) least-squares principle*. Of course, other choices also exists: we could, e.g., minimize the absolute sum of the residuals,

$$\sum_{i=1}^{N} |r_i|,$$

or we could minimize the largest absolute residual:

$$\max_i(|r_i|).$$

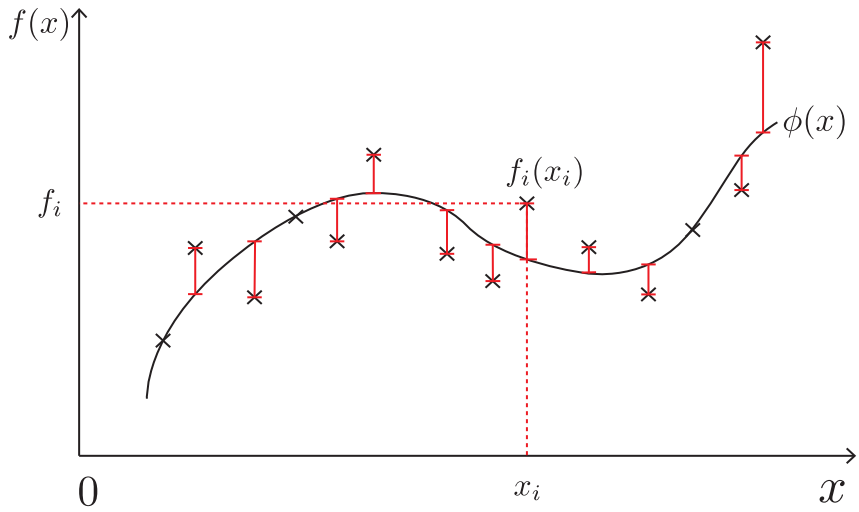

Figure 5.1: Example of regression of samples of a function $f(x)$ with the curve $\phi(x)$. Red bars show the *residuals*, which are minimized in order to solve for ϕ.

The advantage of the least-squares principle is that it is the only one among the three principles, which yields a *linear* system of equations for the unknown coefficients a_0 and a_1. That is the main reason why this principle has become so popular. Let us determine the coefficients a_0 and a_1 according to the least-squares principle. We define a function Φ, which is equal to the square sum of the residuals:

$$\Phi(a_0, a_1) := \sum_{i=1}^{3} r_i^2 = \sum_{i=1}^{3} (f_i - a_0 - a_1\, x_i)^2.$$

We have written $\Phi(a_0, a_1)$ to emphasize that the square sum of the residuals is seen as a function of the unknown coefficients a_0 and a_1. Hence, minimizing the square sum of the residuals means to look for the minimum of the function $\Phi(a_0, a_1)$. A necessary condition for Φ to attain a minimum is that the first derivatives with respect to a_0 and a_1 are equal to zero:

$$\frac{\partial \Phi}{\partial a_0} = -2 \sum_{i=1}^{3} (f_i - a_0 - a_1\, x_i) = 0,$$
$$\frac{\partial \Phi}{\partial a_1} = -2 \sum_{i=1}^{3} (f_i - a_0 - a_1\, x_i) x_i = 0.$$

This is a system of 2 equations for the 2 unknowns a_0 and a_1. It is called *normal equations*. The solution of the normal equations is sometimes called the *least-squares solution*, denoted $\hat{a}_0$ and $\hat{a}_1$. This is mostly done to emphasize that other solutions are possible, as well, as

outlined before. Adopting this notation for the least-squares solution, the normal equations are written as

$$\sum_{i=1}^{3}(\hat{a}_0 + \hat{a}_1\, x_i) = \sum_{i=1}^{3} f_i$$
$$\sum_{i=1}^{3}(\hat{a}_0\, x_i + \hat{a}_1\, x_i^2) = \sum_{i=1}^{3} f_i\, x_i,$$

and in matrix-vector notation

$$\begin{pmatrix} \sum_{i=1}^{3} 1 & \sum_{i=1}^{3} x_i \\ \sum_{i=1}^{3} x_i & \sum_{i=1}^{3} x_i^2 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{3} f_i \\ \sum_{i=1}^{3} f_i\, x_i \end{pmatrix}. \tag{5.3}$$

Numerically, we find

$$\begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} 9.5 \\ 7 \end{pmatrix}.$$

The solution is (to 4 decimal places) $\hat{a}_0 = 4.4167$ and $\hat{a}_1 = -1.2500$. Hence, the least-squares approximation of the given data by a straight line is

$$\hat{\phi}(x) = 4.4167 - 1.2500\, x.$$

The least-squares residuals are computed as $\hat{r}_i = f_i - \hat{\phi}_i$, which gives $\hat{r}_1 = 0.0833$, $\hat{r}_2 = -0.1667$, and $\hat{r}_3 = 0.0833$. The square sum of the (least-squares) residuals is $\hat{\Phi} = \sum_{i=1}^{3} \hat{r}_i^2 = 0.0417$. Notice that the numerical values of the coefficients found before is the choice which yields the smallest possible square sum of the residuals. No other pair of coefficients a_0, a_1 yields a smaller Φ (try it yourself!).

In order to give the normal equations more 'structure', we can define the following *scalar product* of two functions given on a set of N points x_i:

$$\langle f, g\rangle := \sum_{i=1}^{N} f(x_i) g(x_i).$$

Obviously, $\langle f, g\rangle = \langle g, f\rangle$, i.e., the scalar product is symmetric. Using this scalar product, the normal equations (5.3) can be written as (try it yourself!):

$$\begin{pmatrix} < 1, 1 > & < 1, x > \\ < 1, x > & < x, x > \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} < f, 1 > \\ < f, x > \end{pmatrix}.$$

5.3 Least-squares approximation - The general case

Let us now generalize this approach to an arbitrary number N of given data and an approximation function of type

$$\phi(\mathbf{x}) = \sum_{i=1}^{M} a_i\, \varphi_i(\mathbf{x}), \quad M \leq N,$$

for given basis functions $\{\varphi_i(\mathbf{x})\}$. Note that we allow for both univariate and bivariate data. In the univariate case, the location of a data point is uniquely described by 1 variable, denoted e.g., x; in the bivariate case, we need 2 variables to uniquely describe the location of a data point, e.g., the Cartesian coordinates (x, y). Please also notice that the number of basis functions, M, must not exceed the number of data points, N. In least-squares, usually $M \ll N$.
Then, the residuals are (cf. the example above),

$$r(\mathbf{x}_i) = f(\mathbf{x}_i) - \phi(\mathbf{x}_i), \quad i = 1, \ldots, N,$$

or, simply, $r_i = f_i - \phi_i$, and the least-squares principle is

$$\Phi(\mathbf{a}) = \sum_{i=1}^{N} r(\mathbf{x}_i)^2 = \sum_{i=1}^{N} r_i^2 = \langle r, r \rangle \stackrel{!}{=} \text{ minimum}, \tag{5.4}$$

where the vector $\mathbf{a}$ is defined as $\mathbf{a} := (a_1, \ldots, a_M)^T$. The advantage of the use of a *scalar product* becomes clear now, because the normal equations, which are the solution of the minimization problem (5.4), are

$$\begin{pmatrix} \langle \varphi_1, \varphi_1 \rangle & \cdots & \langle \varphi_1, \varphi_M \rangle \\ \vdots & & \vdots \\ \langle \varphi_M, \varphi_1 \rangle & \cdots & \langle \varphi_M, \varphi_M \rangle \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_M \end{pmatrix} = \begin{pmatrix} \langle f, \varphi_1 \rangle \\ \vdots \\ \langle f, \varphi_M \rangle \end{pmatrix}, \tag{5.5}$$

with

$$\langle \varphi_j, \varphi_k \rangle = \sum_{i=1}^{N} \varphi_j(\mathbf{x}_i) \varphi_k(\mathbf{x}_i). \tag{5.6}$$

Note that the normal equations are symmetric, because

$$\langle \varphi_k, \varphi_j \rangle = \langle \varphi_j, \varphi_k \rangle.$$

The solution of the normal equations yields the least-squares estimate of the coefficients $\mathbf{a}$, denoted $\hat{\mathbf{a}}$, and the discrete least-squares approximation is the function

$$\hat{\phi}(\mathbf{x}) = \sum_{i=1}^{M} \hat{a}_i \, \varphi_i(\mathbf{x}).$$

The smallness of the square sum of the residuals,

$$\langle \hat{r}, \hat{r} \rangle = \langle f - \hat{\phi}, f - \hat{\phi} \rangle$$

can be used as a criterion for the efficiency of the approximation. Alternatively, the so-called *root-mean-square (RMS) error* in the approximation is also used as a measure of the fit of the function $\hat{\phi}(\mathbf{x})$ to the given data. It is defined by

$$\sigma_{RMS} := \sqrt{\frac{\langle \hat{r}, \hat{r} \rangle}{N}},$$

where $\hat{r} = f - \hat{\phi}$, and

$$\langle \hat{r}, \hat{r} \rangle = \sum_{i=1}^{N} \Big(f(\mathbf{x}_i) - \hat{\phi}(\mathbf{x}_i) \Big)^2.$$

When *univariate algebraic polynomials* are used as basis functions it can be shown that $N \geq M$ together with the linear independence of the basis functions guarantee the *unique solvability* of the normal equations. For other types of univariate basis functions this is not guaranteed. For $N < M$ the uniqueness gets lost. For $N = M$ we have *interpolation* and $\hat{\phi}(x_i) = f(x_i)$ for $i = 1, \ldots, N$.

For the bivariate case using radial functions as basis functions we have the additional problem that we have less basis functions (namely M) than we have data sites (namely N). Hence, we cannot place below every data point a basis function. Therefore, we need to have a strategy of where to locate the radial functions. This already indicates that the question whether the normal equations can be solved in the bivariate case with radial functions is non-trivial. In fact, we can only guarantee unique solvability of the normal equations for certain radial functions (i.e., multiquadrics, Gaussian) and severe restrictions to the location of the centres of the radial functions (they must be sufficiently well distributed over D in some sense) and to the location of the data sites (they must be fairly evenly clustered about the centres of the radial functions with the diameter of the clusters being relatively small compared to the separation distance of the data sites). Least-squares approximation with radial basis functions is not subject of this course.

The method of (discrete) least squares has been developed by Gauss in 1794 for smoothing data in connection with geodetic and astronomical problems.

Example 5.1
Given 5 function values $f(x_i)$, $i = 1, \ldots, 5$, of the function $f(x) = (1 + x^2)^{-1}$ (see the table below). We look for the discrete least-squares approximation $\hat{\phi}$ among all quadratic polynomials ϕ.

Step 1: *the choice of the basis functions is prescribed by the task description:*

$$\varphi_1 = 1,\ \varphi_2 = x,\ \varphi_3 = x^2 \Rightarrow \phi(x) = \sum_{i=1}^{3} c_i \varphi_i(x).$$

Step 2: *because no other information is available, we choose $w_i = 1$, $i = 1, \ldots, 5$.*

Step 3: *the given values $f(x_i)$, $i = 1, \ldots, 5$*

i	1	2	3	4	5
x_i	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1
$f(x_i)$	0.5	0.8	1	0.8	0.5

lead to the normal equations

$$\begin{pmatrix} 5 & 0 & 2.5 \\ 0 & 2.5 & 0 \\ 2.5 & 0 & 2.125 \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \hat{c}_3 \end{pmatrix} = \begin{pmatrix} 3.6 \\ 0 \\ 1.4 \end{pmatrix}$$

with the solution (to 5 decimal places) $\hat{c}_1 = 0.94857$, $\hat{c}_2 = 0.00000$, $\hat{c}_3 = -0.45714$, *yielding* $\hat{\phi}(x) = 0.94857 - 0.45714\,x^2, \quad x \in [-1, 1]$. *Under all quadratic polynomials,* $\hat{\phi}(x)$ *is the best approximation of* f *in the discrete least squares sense. For* $x = 0.8$, *we obtain for instance* $\hat{\phi}(0.8) = 0.65600$. *The absolute error at* $x = 0.8$ *is* $|f(0.8) - \hat{\phi}(0.8)| = 4.6 \cdot 10^{-2}$. *The results are plotted in figure 5.2.*

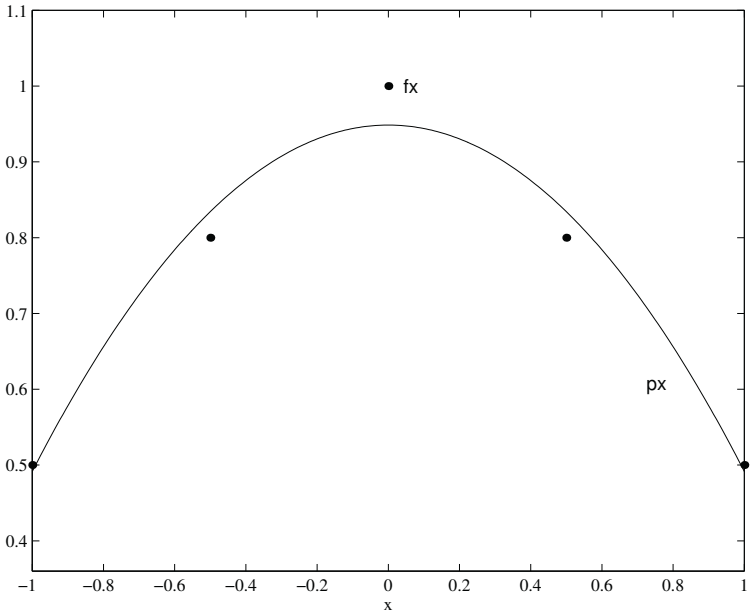

Figure 5.2: The 5 data points given by $f(x) = (1+x^2)^{-1}$ (dots) and the best approximation ϕ (solid line) of all quadratic polynomials in the discrete least squares sense.

Exercise 5.2

Given points $(x_i, f(x_i))$, $i = 1, \ldots, 4$.

i	1	2	3	4
x_i	0.02	0.10	0.50	1.00
$f(x_i)$	50	10	1	0

We look for the best approximation $\hat{\phi}$ *in the discrete least squares sense of all functions* $\phi(x) = \sum c_k \varphi_k(x)$ *and the following basis functions:*

1. $\varphi_1 = 1$, $\varphi_2 = x$
2. $\varphi_1 = 1$, $\varphi_2 = x$, $\varphi_3 = x^2$
3. $\varphi_1 = 1$, $\varphi_2 = x$, $\varphi_3 = x^2$, $\varphi_4 = x^3$
4. $\varphi_1 = 1$, $\varphi_2 = 1/x$

Give a graphical representation of the four different $\hat{\phi}$. What choice of the basis functions yields the "best" result?

5.4 Weighted least-squares (⋆⋆ not examined)

We can slightly generalize the least-squares method by assigning to each data value f_i a so-called *weight* $w_i > 0$. This may be justified, for instance, if the accuracy of the data values vary, i.e., if one data point, say, the one with index i, is more accurate than another one, say, with index j. If this is true, then it is natural to expect that $|f_i - \phi_i|$ is smaller than $|f_j - \phi_j|$. To achieve this, we have to assign the data point with index i a larger weight than the data point with index j. The corresponding method is called *weighted least-squares method*. It can be shown that the normal equations associated with the weighted least-squares method are formally identical to the normal equations associated with the classical least-squares method (which uses unit weights $w_i = 1$) if we slightly redefine the scalar product: instead of the definition (5.6), we use

$$\langle \varphi_j, \varphi_k \rangle := \sum_{i=1}^{N} w_i \, \varphi_j(\mathbf{x}_i) \varphi_k(\mathbf{x}_i). \tag{5.7}$$

The *weighted least-squares principle* is

$$\Phi(\mathbf{a}) = \langle r, r \rangle = \sum_{i=1}^{N} w_i (f_i - \phi_i)^2 \stackrel{!}{=} \text{ minimum}, \tag{5.8}$$

and the weighted least-squares solution is given by Eq. (5.5) when using the definition of the scalar product, Eq. (5.7). Note that according to Eq. (5.8), the weighted least-squares method does not minimize the square sum of the residuals, but the weighted square sum of the residuals. The RMS error in the weighted least-squares approximation is

$$\sigma_{RMS} = \sqrt{\frac{\langle \hat{r}, \hat{r} \rangle}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} w_i \hat{r}_i^2}.$$