

## NOTES ON LEAST SQUARES APPROXIMATION

Given  $n$  data points  $(x_1, y_1), \dots, (x_n, y_n)$ , we would like to find the line  $L$ , with an equation of the form  $y = mx + b$ , which is the “best fit” for the given data points. We will do this using orthogonal projections and a general approximation theorem from linear algebra, which we now recall.

## 1 Orthogonal projections and the approximation theorem

Let  $V$  be a Euclidean space, and let  $W$  be a finite-dimensional subspace of  $V$ . Choose an orthogonal basis  $\{e_1, \dots, e_m\}$  for  $W$  (which exists by the Gram-Schmidt procedure).

**Definition 1.** The *orthogonal projection* of an element  $x \in V$  onto  $W$  is given by the formula

$$p_W(x) = \sum_{i=1}^m \frac{\langle x, e_i \rangle}{\langle e_i, e_i \rangle} e_i.$$

It is not obvious from the definition that  $p_W(x)$  is independent of the choice of an orthogonal basis for  $W$ , but this is true, and follows from the approximation theorem below.

We will use the following fact, which lies at the heart of the Gram-Schmidt procedure, and which justifies the name “orthogonal projection”:

**Lemma 1.** For every  $x \in V$ , we have  $x - p_W(x) \in W^\perp$ .

*Proof.* It suffices to show that  $\langle x - p_W(x), e_j \rangle = 0$  for each  $j = 1, \dots, m$ . This follows from the following computation:

$$\begin{aligned} \langle x - p_W(x), e_j \rangle &= \langle x, e_j \rangle - \langle p_W(x), e_j \rangle \\ &= \langle x, e_j \rangle - \left\langle \sum_{i=1}^m \frac{\langle x, e_i \rangle}{\langle e_i, e_i \rangle} e_i, e_j \right\rangle \\ &= \langle x, e_j \rangle - \frac{\langle x, e_j \rangle}{\langle e_j, e_j \rangle} \langle e_j, e_j \rangle \\ &= 0. \end{aligned}$$

□

We can now state the main result of this section:

**Theorem 1 (The approximation theorem).** *The orthogonal projection  $p_W(x)$  is closer to  $x$  than any other element of  $W$ .*

*Proof.* For any  $y \in W$ , we can write  $x - y = (x - p_W(x)) + (p_W(x) - y)$ . We have  $p_W(x) - y \in W$ , and by the above lemma we know that  $x - p_W(x) \in W^\perp$ . The Pythagorean theorem (for general Euclidean spaces) now shows that

$$\begin{aligned}\|x - y\|^2 &= \|x - p_W(x)\|^2 + \|p_W(x) - y\|^2 \\ &\geq \|x - p_W(x)\|^2\end{aligned}$$

with equality if and only if  $y = p_W(x)$ . □

Note that it follows from the approximation theorem that  $p_W(x)$  is independent of the choice of an orthogonal basis for  $W$ , since we have characterized  $p_W(x)$  by a condition which does not make reference to any particular basis.

## 2 The nearest solution to an overdetermined system

A problem which arises in many contexts, including least squares approximation, is the following. Suppose  $A$  is an  $m \times n$  matrix with more rows than columns, and that the rank of  $A$  equals the number of columns. If a vector  $y \in \mathbf{R}^m$  is not in the image of  $A$ , then (by definition) the equation  $Ax = y$  has no solution. In practice, one often wants to find a “best approximate solution” to such a system, i.e., a vector  $x \in \mathbf{R}^n$  for which  $\|Ax - y\|$  (or equivalently,  $\|Ax - y\|^2$ ) is as small as possible.

To do this, we recall that the column space  $C$  of  $A$  coincides with the image (= range) of  $A$ . (This follows easily from the fact that if  $A_1, \dots, A_m$  are the columns of  $A$  and  $e_1, \dots, e_m$  are the standard unit coordinate vectors in  $\mathbf{R}^m$ , then  $Ae_i = A_i$ .) Recall also that since  $A$  is assumed to have rank  $n$ , the kernel of  $A$  equals  $\{0\}$ .

**Theorem 2.** *Let  $A$  be an  $m \times n$  matrix with rank  $n$ , and let  $P = P_C$  denote orthogonal projection onto the image of  $A$ . Then for every  $y \in \mathbf{R}^m$ , the*

equation  $Ax = Py$  has a unique solution  $x_* \in \mathbf{R}^n$ . Moreover,  $x_*$  is the best approximate solution to the equation  $Ax = y$ , in the sense that for any  $x \in \mathbf{R}^n$ ,

$$\|Ax_* - y\|^2 \leq \|Ax - y\|^2$$

with equality if and only if  $x = x_*$ .

*Proof.* By definition, the orthogonal projection  $Py$  belongs to the image of  $A$ . Therefore  $Ax_* = Py$  for some  $x_* \in \mathbf{R}^n$ . Moreover,  $x_*$  is uniquely determined, since if  $Ax_1 = Ax_2$  then  $A(x_1 - x_2) = 0$  and  $x_1 - x_2 \in \ker(A)$ . But  $\ker(A) = \{0\}$  by hypothesis (since  $A$  has rank  $n$ ), so  $x_1 - x_2 = 0$ , i.e.,  $x_1 = x_2$ .

By the approximation theorem, we know that

$$\|Py - y\|^2 \leq \|Ax - y\|^2$$

for every  $x \in \mathbf{R}^n$ , with equality if and only if  $Ax = Py$ . Substituting  $Ax_*$  for  $Py$  into this inequality gives the desired result.  $\square$

### 3 Least squares approximation

We now return to the least squares approximation problem. Given  $n$  data points  $(x_1, y_1), \dots, (x_n, y_n)$  in  $\mathbf{R}^2$ , we would like to find a line  $L$  of the form  $y = mx + b$  which is the “closest fit” for the given data points, in the sense that the “least squares error” term

$$S(m, b) = \sum_{i=1}^n (mx_i + b - y_i)^2$$

is as small as possible. A method for doing this was first developed by Legendre and Gauss between 1805 and 1810 in connection with astronomical observations.

To find a formula for the “least squares regression line”  $L$ , we note that the system of  $n$  equations in the two unknowns  $m$  and  $b$

$$\begin{aligned} mx_1 + b &= y_1 \\ &\vdots \\ mx_n + b &= y_n \end{aligned}$$

is overdetermined. If we assume that at least two of the the  $x$ -coordinates  $x_1, \dots, x_n$  are distinct, then the matrix

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

has rank 2, and the system we are trying to solve can be written as  $Av = Y$ , where

$$v = \begin{bmatrix} m \\ b \end{bmatrix}, X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Note that there is no reason that  $Y$  should lie in the image of  $A$ , so this system typically has no solution. However, let  $P$  denote orthogonal projection onto the column space (= image) of  $A$ . Then by Theorem 2, there is a unique solution  $v_* = (m_*, b_*)$  to the equation  $Av = PY$ , and this solution minimizes the quantity  $\|Av - Y\|^2$ . Since  $\|Av - Y\|^2 = S(m, b)$ , it follows that the best-fit line  $L$  that we are looking for is precisely the line given by the equation  $y = m_*x + b_*$ .

We will now derive a concrete formula for  $m_*$  and  $b_*$ , and hence for the least squares regression line  $L$ . The idea is to use the Gram-Schmidt procedure to find an explicit formula for the orthogonal projection  $PY$ , and then to solve the linear system  $Av = PY$  by elimination of variables.

Let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ , and let

$$\bar{X} = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix}, \bar{Y} = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}.$$

With this notation, we have:

**Theorem 3.** *The least squares regression line  $L$  is given by the equation  $y = m_*x + b_*$ , where*

$$m_* = \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{(X - \bar{X}) \cdot (X - \bar{X})}$$

and

$$b_* = \bar{y} - m_*\bar{x}.$$

*Proof.* By Gram-Schmidt, an orthogonal basis for the column space  $C$  of  $A$  is given by

$$v_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \frac{\sum x_i}{n} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = X - \bar{X}.$$

Therefore the orthogonal projection  $PY$  of  $Y$  onto  $C$  is given by

$$\begin{aligned} PY &= \frac{Y \cdot v_1}{v_1 \cdot v_1} v_1 + \frac{Y \cdot v_2}{v_2 \cdot v_2} v_2 \\ &= \bar{Y} + a(X - \bar{X}), \end{aligned}$$

where

$$a = \frac{Y \cdot (X - \bar{X})}{(X - \bar{X}) \cdot (X - \bar{X})}.$$

Since the coordinates of  $X - \bar{X}$  sum to zero, it follows easily that  $\bar{Y} \cdot (X - \bar{X}) = 0$ , and therefore

$$a = \frac{(Y - \bar{Y}) \cdot (X - \bar{X})}{(X - \bar{X}) \cdot (X - \bar{X})}.$$

Finally, we solve the equation  $Av = PY$ . Suppose that  $x_i \neq x_j$ . Considering the  $i$ th and  $j$ th rows of  $A$ , we find the equations

$$\begin{aligned} m_* x_i + b_* &= \bar{y} + a(x_i - \bar{x}) \\ m_* x_j + b_* &= \bar{y} + a(x_j - \bar{x}) \end{aligned}$$

and solving this system yields (after a little bit of computation)

$$m_* = a, b_* = \bar{y} - a\bar{x}$$

as desired. □

**Example 1.** Suppose the three data points are  $(1, 2), (2, 5), (3, 7)$ . Then  $\bar{x} = 2$  and  $\bar{y} = 14/3$ . We have  $X = (1, 2, 3)$  and  $Y = (2, 5, 7)$ , so that  $X - \bar{X} = (-1, 0, 1)$  and  $Y - \bar{Y} = (-8/3, 1/3, 7/3)$ . Therefore  $m_* = 5/2$  and  $b_* = -1/3$ , so that the least squares regression line is given by the equation  $y = \frac{5}{2}x - \frac{1}{3}$ .