EE951: Linear Algebra **eMasters**

Chapter 10: Linear Systems

Author: Ketan Rajawat

Contents

10.1 Introduction

A system of linear equations or a linear system is a collection of one or more linear equations involving the same variables. Using the notation introduced thus far and for $\mathbf{x} \in \mathbb{R}^n$, we can write m equations

$$\mathbf{a}_1^\mathsf{T}\mathbf{x} = b_1 \tag{10.1}$$
$$\mathbf{a}_2^\mathsf{T}\mathbf{x} = b_2 \tag{10.2}$$
$$\vdots \tag{10.3}$$
$$\mathbf{a}_m^\mathsf{T}\mathbf{x} = b_m \tag{10.4}$$

compactly as

$$\mathbf{Ax} = \mathbf{b} \tag{10.5}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T} \\ \mathbf{a}_2^\mathsf{T} \\ \vdots \\ \mathbf{a}_m^\mathsf{T} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{10.6}$$

Studying and solving such systems of linear equations can be considered as the primary goal of linear algebra. Indeed, our focus so far has been on introducing the concepts and notation surrounding linear systems. In this

chapter, we will define and mention the simplest cases when such systems can be solved easily. Subsequent chapters will look at more sophisticated cases and the different computational algorithms for solving them.

We will discuss two specific cases in this chapter. The first case is when $\mathbf{A}$ is square and invertible, so that the system of equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution. The second case is when $\mathbf{A}$ is tall and the system of equations $\mathbf{Ax} = \mathbf{b}$ has no solution, so that we need to find $\mathbf{x}$ such that $\mathbf{Ax} \approx \mathbf{b}$.

10.2 Matrix Inverse

Recall that we designate $x = 1/a = a^{-1}$ as the multiplicative inverse of a because $xa = 1$. The inverse exists only when a is not equal to zero. The notion of the multiplicative inverse can similarly be explored for matrices. However, the extension is not a simple matter. The conditions under which inverse may exist are complicated and deferred to a later point.

Formally, for a square matrix $\mathbf{A}$, if there exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, then $\mathbf{A}$ is said to be *invertible* or *non-singular*. If no such matrix $\mathbf{A}^{-1}$ exists, then $\mathbf{A}$ is called *non-invertible* or *singular*.

It is easy to verify that $\mathbf{I}^{-1} = \mathbf{I}$. The inverse of a matrix $\mathbf{A}$ has several important properties:

1. *Uniqueness:* If $\mathbf{A}$ is invertible, then its inverse $\mathbf{A}^{-1}$ is unique. There is only one matrix that can serve as the inverse of $\mathbf{A}$. If $\mathbf{B}$ and $\mathbf{C}$ are two different inverses of $\mathbf{A}$, then we can see that $\mathbf{B} = \mathbf{BI} = \mathbf{B}(\mathbf{AC}) = (\mathbf{BA})\mathbf{C} = \mathbf{IC} = \mathbf{C}$ since $\mathbf{AC} = \mathbf{BA} = \mathbf{I}$.
2. *Inverse of product:* For invertible matrices $\mathbf{A}$, $\mathbf{B}$, it holds that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. To see this, let $\mathbf{X} = (\mathbf{AB})^{-1}$ so that

   $$\mathbf{ABX} = \mathbf{I} \tag{10.7}$$

   Then multiplying both sides by $\mathbf{A}^{-1}$ on the left, we obtain

   $$\mathbf{BX} = \mathbf{A}^{-1} \tag{10.8}$$

   Subsequently, multiplying both sides by $\mathbf{B}^{-1}$ on the left, we obtain

   $$\mathbf{X} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{10.9}$$

3. *Inverse of the Inverse:* If $\mathbf{A}$ is invertible, then $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
4. *Inverse of transpose:* If $\mathbf{A}$ is invertible, then $(\mathbf{A}^{-1})^{\mathsf{T}} = (\mathbf{A}^{\mathsf{T}})^{-1}$.

As an example, consider the diagonal matrix, where all the elements outside the main diagonal are zeros. If $\mathbf{D}$ is a non-singular diagonal matrix, we can find its inverse by taking the reciprocal of each non-zero diagonal element. For example, if $\mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, then $\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$. Consider a general 2x2 matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $ad - bc \neq 0$ to ensure that it is invertible. The inverse of $\mathbf{A}$ is given by:

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

For example, if $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$, then $\mathbf{A}^{-1} = \frac{1}{-2} \begin{bmatrix} 5 & -3 \\ -4 & 2 \end{bmatrix} = \begin{bmatrix} -\frac{5}{2} & \frac{3}{2} \\ 2 & -1 \end{bmatrix}$. We observe here that for a matrix, if $ad = bc$, the denominator becomes zero and hence the inverse cannot be calculated. Such a matrix would be non-invertible.

10.2.1 Solving $\mathbf{Ax} = \mathbf{b}$

For a square system, i.e., when $\mathbf{A} \in \mathbb{R}^{n \times n}$, and when $\mathbf{A}$ is invertible, the system of equations $\mathbf{Ax} = \mathbf{b}$ can be solved by pre-multiplying both sides by $\mathbf{A}^{-1}$ to yield

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{10.10}$$

Since $\mathbf{A}^{-1}$ is unique, the solution to such a system of equations is also unique.

10.3 Orthogonal Matrices

A square matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is said to be orthogonal if it satisfies $\mathbf{QQ}^\mathsf{T} = \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}$. The inverse of an orthogonal matrix $\mathbf{Q}$ is its transpose, i.e., $\mathbf{Q}^{-1} = \mathbf{Q}^\mathsf{T}$. If the vector represents displacement, multiplication with an orthogonal matrix amounts to keeping the length of the displacement same but rotating it by a certain angle. As an example, suppose we have a point (x, y) in the Cartesian coordinate system. To rotate this point counterclockwise by an angle θ, we can use an orthogonal matrix. The general form of a 2D rotation matrix, denoted as $\mathbf{R}$, is:

$$\mathbf{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

We can verify that $\mathbf{R}$ is an orthogonal matrix since

$$\mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{R}\mathbf{R}^\mathsf{T} = \begin{bmatrix} \cos^2(\theta) + \sin^2(\theta) & \cos(\theta)\sin(\theta) - \sin(\theta)\cos(\theta) \\ \sin(\theta)\cos(\theta) - \cos(\theta)\sin(\theta) & \sin^2(\theta) + \cos^2(\theta) \end{bmatrix} = \mathbf{I}$$

To apply this rotation to the point (x, y), we can represent the point as a column vector $\mathbf{p} = \begin{bmatrix} x \\ y \end{bmatrix}$. The rotated point $\mathbf{p}'$ can be obtained by multiplying the rotation matrix $\mathbf{R}$ with $\mathbf{p}$ as follows:

$$\mathbf{p}' = \mathbf{Rp}$$

Let's illustrate this with an example. Suppose we want to rotate the point $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ counterclockwise by an angle of 45°. The rotation matrix for this angle is:

$$\mathbf{R} = \begin{bmatrix} \cos(45°) & -\sin(45°) \\ \sin(45°) & \cos(45°) \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

To find the rotated point $\mathbf{p}'$, we multiply $\mathbf{R}$ by $\mathbf{p}$:

$$\mathbf{p}' = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Therefore, the point $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ rotated counterclockwise with respect to the origin by 45° becomes $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, as also depicted in Fig. 10.1. Consider a system of equations of the form $\mathbf{Qx} = \mathbf{b}$ where $\mathbf{Q}$ is orthogonal. Then the solution is given by $\mathbf{Q}^\mathsf{T}\mathbf{b}$, which requires $2n^2 - n$ flops for the matrix-vector product, as opposed to $\frac{2}{3}n^3$ flops required in the general case.

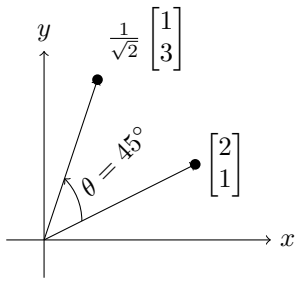

Figure 10.1: Point $(2, 1)$ rotated counterclockwise with respect to the origin by 45°

10.3.1 Orthonormal Vectors

Recall that an orthogonal matrix $\mathbf{Q}$ is a square matrix that satisfies

$$\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\mathsf{T} = \mathbf{I} \tag{10.11}$$

Lemma 10.1. The columns of an orthogonal matrix are orthonormal.

Proof: Consider an orthogonal matrix $\mathbf{Q}$. By definition, an orthogonal matrix satisfies $\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}$. Let the columns of $\mathbf{Q}$ be denoted as $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$. From the orthogonality of $\mathbf{Q}$ and recalling the result for Gram matrices:

$$(\mathbf{Q}^\mathsf{T}\mathbf{Q})_{ij} = \langle \mathbf{q}_i, \mathbf{q}_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \tag{10.12}$$

Therefore, the columns of $\mathbf{Q}$ form an orthonormal set. ■

The converse is not necessarily true. Consider n orthonormal vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and let

$$\mathbf{A} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_n \end{bmatrix} \tag{10.13}$$

Then from Gram-matrix representation, it is easy to see that $\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$. But if $\mathbf{v}_i \in \mathbb{R}^m$, then $\mathbf{A} \in \mathbb{R}^{m \times n}$ is not a square matrix. We will show later that if $m = n$, then $\mathbf{A}$ is orthogonal but not otherwise.

Example 10.1. Consider

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{10.14}$$

Then it is easy to see that

$$\mathbf{A}^\mathsf{T}\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{10.15}$$

but

$$\mathbf{A}\mathbf{A}^\mathsf{T} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{10.16}$$

Orthogonal matrices are length and inner-product preserving, in the following sense:

$$\langle \mathbf{Q}\mathbf{u}, \mathbf{Q}\mathbf{v} \rangle = (\mathbf{Q}\mathbf{u})^\mathsf{T}\mathbf{Q}\mathbf{v} = \mathbf{u}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{Q}\mathbf{v} = \mathbf{u}^\mathsf{T}\mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle \tag{10.17}$$

$$\|\mathbf{Q}\mathbf{u}\|^2 = \langle \mathbf{Q}\mathbf{u}, \mathbf{Q}\mathbf{u} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle = \|\mathbf{u}\|^2 \tag{10.18}$$

10.4 Gradients and Minimization

Let us recall the derivative condition for minimization of scalar-valued function $h : \mathbb{R} \to \mathbb{R}$. If $\mathbf{x}^\star$ minimizes $h(x)$, its derivative must be zero at $x^\star$, i.e.,

$$\frac{dh(x)}{dx} \mid_{x=x^\star} = 0$$

In other words, if our goal is to find the $x^\star$ that minimizes $h(x)$, we can find the solution to the equation in (10.4). For instance, if $h(x) = \log(x) - ax$ for $a > 0$, its derivative is given by $1/x - a$. Therefore, the derivative becomes zero when $x^\star = 1/a$.

The approach also generalizes to functions $h : \mathbb{R}^n \to \mathbb{R}$. If the vector $\mathbf{x}^\star$ minimizes $h(\mathbf{x})$, then it must also satisfy

$$\frac{\partial h(\mathbf{x})}{\partial x_i} \mid_{\mathbf{x}=\mathbf{x}^\star} = 0 \qquad\qquad i = 1, 2, \ldots, n \tag{10.19}$$

The equation can be compactly written in vector notation as $\nabla h(\mathbf{x}^\star) = 0$, where the gradient vector $\nabla h(\mathbf{x}) \in \mathbb{R}^n$ has entries:

$$\big[\nabla h(\mathbf{x})\big]_i = \frac{\partial h(\mathbf{x})}{\partial x_i} \qquad\qquad i = 1, 2, \ldots, n \tag{10.20}$$

As an example, consider the function .

$$h(\mathbf{x}) = \|\mathbf{x} - \mathbf{b}\|^2 = \sum_{i=1}^{n} (x_i - b_i)^2 \tag{10.21}$$

so that

$$\frac{\partial h(\mathbf{x})}{\partial x_i} = 2(x_i - b_i) \tag{10.22}$$

and hence the partial derivatives vanish when $x_i = b_i$ for all i, or equivalently $\mathbf{x}^\star = \mathbf{b}$.

10.5 Least Squares

The system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ for $m > n$ and $\mathbf{A}^\mathsf{T}\mathbf{A}$ is invertible, is said to be over-determined since there are more equations than the number of variables. In general, such a system will not have a solution, i.e., there is no vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{b}$. The least-square problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{10.23}$$

represents a possible compromise: find the vector $\mathbf{x}$ that ensures that the total squared error $\|\mathbf{Ax}-\mathbf{b}\|^2$ is minimized. If we are able to find a solution $\mathbf{x}^\star$ for which the error is very small, it follows that $\mathbf{Ax}\approx\mathbf{b}$. The following lemma provides the solution of the least-squares problem for a specific case.

Lemma 10.2. If $\mathbf{A}^\mathsf{T}\mathbf{A}$ is invertible, the solution of the least-squares problem is given by

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|^2 = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} \tag{10.24}$$

Before looking at the general proof, let us consider the simpler case of $n=1$.

Proof for $n=1$: We will establish a general proof and develop more intuition later. However, let us see an example for the case of $n=1$. That is, we want to solve

$$\mathbf{a}x = \mathbf{b} \tag{10.25}$$

where $\mathbf{a}\in\mathbb{R}^m$ while $x\in\mathbb{R}$. We note that this equation generally has no solution, unless $b_i/a_i = b_j/a_j$ for all $1\le i\ne j\le m$. The least square solution

$$x^\star = \arg\min_x \|\mathbf{a}x-\mathbf{b}\|^2 = \arg\min_x \sum_{i=1}^m (a_ix-b_i)^2. \tag{10.26}$$

We can calculate the minimum by differentiating with respect to x and then setting the derivative to zero, which yields:

$$\frac{d}{dx}\sum_{i=1}^m (a_ix-b_i)^2 = 2\sum_{i=1}^m a_i(a_ix-b_i) = 0 \tag{10.27}$$

$$\Leftrightarrow x^\star = \frac{\sum_{i=1}^m a_ib_i}{\sum_{i=1}^m a_i^2} = \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\mathbf{a}^\mathsf{T}\mathbf{a}} \tag{10.28}$$

■

Proof for general n: We now establish the proof for the general case. We begin by calculating the gradient of the function

$$h(\mathbf{x}) = \|\mathbf{Ax}-\mathbf{b}\|^2 = \sum_{i=1}^m (\mathbf{a}_{i\cdot}^\mathsf{T}\mathbf{x}-b_i)^2 \tag{10.29}$$

$$= \sum_{i=1}^m \left(\sum_{k=1}^n A_{ik}x_k - b_i\right)^2 \tag{10.30}$$

The partial derivative of h with respect to x_j is given by

$$\frac{\partial h(\mathbf{x})}{\partial x_j} = 2\sum_{i=1}^m \left(\sum_{k=1}^n A_{ik}x_k - b_i\right) A_{ij} \tag{10.31}$$

$$= 2\sum_{i=1}^m\sum_{k=1}^n A_{ij}A_{ik}x_k - 2\sum_{i=1}^m A_{ij}b_i \tag{10.32}$$

$$= 2\sum_{k=1}^n \left(\sum_{i=1}^m A_{ij}A_{ik}\right)x_k - 2\sum_{i=1}^m A_{ij}b_i \tag{10.33}$$

Consider the matrix $\mathbf{B}$ whose (j,k)-th entry is given by

$$B_{jk} = \sum_{i=1}^m A_{ij}A_{ik} = \left[\mathbf{A}^\mathsf{T}\mathbf{A}\right]_{jk} \tag{10.34}$$

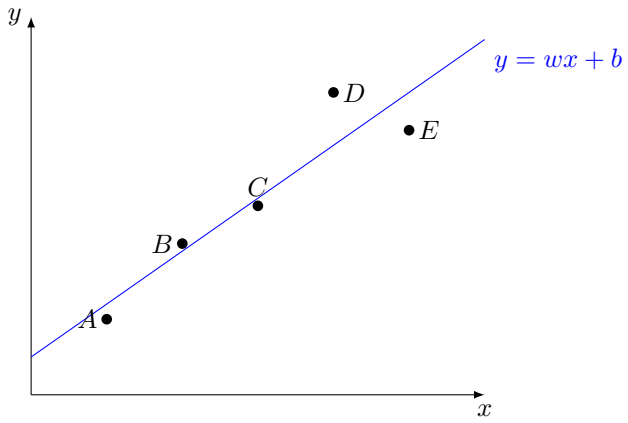

Figure 10.2: Fitting a line in 2D

which implies that $\mathbf{B} = \mathbf{A}^\mathsf{T}\mathbf{A}$. Likewise, we can see that

$$\left[\mathbf{A}^\mathsf{T}\mathbf{b}\right]_j = \sum_{i=1}^{m} A_{ij} b_i \tag{10.35}$$

Hence, we can say that

$$\nabla h(\mathbf{x}) = \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{A}^\mathsf{T}\mathbf{b} \tag{10.36}$$

Therefore, the zero-gradient condition is given by

$$\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{b} \tag{10.37}$$

This system of equations is also called the *normal equations*. Since we assumed that $(\mathbf{A}^\mathsf{T}\mathbf{A})$ is invertible, it follows immediately that the solution of this system of equations (i.e., the point where the gradient becomes zero) is given by

$$\mathbf{x}^\star = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} \tag{10.38}$$

■

10.5.1 Fitting a line

The problem of fitting a line that passes through a given set of m points can be cast as a least-squares problem. The objective is to find the best-fitting line that minimizes the overall squared distance between the line and the data points. Let us consider a set of m points in a two-dimensional space, denoted as (x_1, y_1), (x_2, y_2), …, (x_m, y_m). We want to find a line that can be represented by the equation $y = wx + b$, where w is the slope and b is the y-intercept. An example is shown in Fig. 10.2.

To determine the best-fitting line, we need to minimize the sum of squared distances between each data point and the line. For a given point (x_i, y_i), the squared distance between the point and the line is given by $(y_i - wx_i - b)^2$. Therefore, we can define the objective function as follows:

$$\min_{w,b} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

To arrange the objective in the canonical form, we need to rewrite the objective function in matrix form. We can define the design matrix $\mathbf{X}$ as:

$$\mathbf{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}$$

Here, each row of $\mathbf{X}$ represents a data point $(x_i, 1)$, and the first column corresponds to the x-coordinates of the points, while the second column is filled with ones. Next, we can define the coefficient vector $\mathbf{u} = \begin{bmatrix} w \\ b \end{bmatrix}$. Similarly, we can define the target vector $\mathbf{y}$ as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

The squared distances between the points and the line can be represented as $\mathbf{e}^T\mathbf{e}$ where

$$\mathbf{e} = \begin{bmatrix} y_1 - wx_1 - b \\ y_2 - wx_2 - b \\ \vdots \\ y_m - wx_m - b \end{bmatrix}.$$

The objective function can be rewritten as the sum of squared distances:

$$\min_{\mathbf{u}} \|\mathbf{e}\|^2 = \min_{\mathbf{u}} \mathbf{e}^T\mathbf{e}$$

Substituting the expressions for $\mathbf{X}$, $\mathbf{u}$, and $\mathbf{y}$, we can rewrite the objective function as:

$$\min_{\mathbf{u}} \|\mathbf{y} - \mathbf{X}\mathbf{u}\|^2$$

Proceeding as earlier, it is possible to verify that

$$w^\star = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2} \tag{10.39}$$

$$b^\star = \bar{y} - w^\star \bar{x} \tag{10.40}$$

where $\bar{x} = \frac{1}{m}\sum_{i=1}^m x_i$ is the mean of the x values, and $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$ is the mean of the y values.

We remark that the least squares formulation also applies to cases when we want to fit a linear combination of (possibly non-linear) functions to a given set of points. In other words, if we hypothesize that a given set of points (x_1, y_1), (x_2, y_2), …, (x_m, y_m) are related as

$$\mathbf{y}_i \approx a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_k f_k(x_i) \tag{10.41}$$

then we can cast the problem of finding $a_1, a_2, \ldots, a_k$ as a least-squares problem. In the simplest case discussed earlier, we have $f_1(x) = x$ and $f_2(x) = 1$. When fitting a parabola to a given set of points, we will

have $f_1(x) = x^2$, $f_2(x) = x$, and $f_3(x) = 1$, so that the system of equations becomes:

$$\begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ & \vdots & \\ x_m^2 & x_m & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{10.42}$$

which can again be approximately solved using the normal equations.