
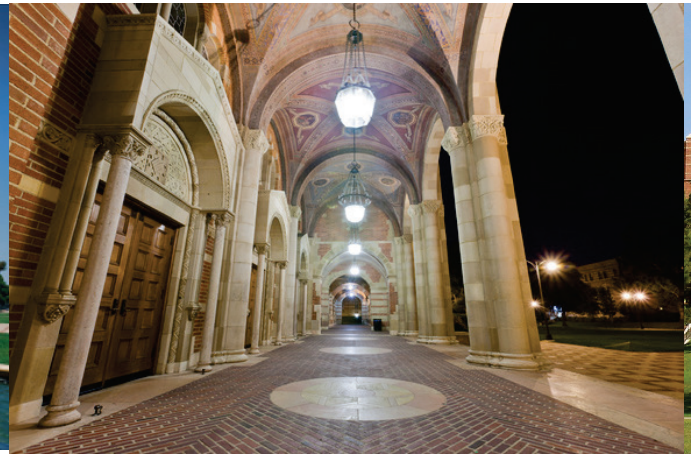


EE210A: Adaptation and Learning

Professor Ali H. Sayed

LECTURE #14

LEAST-SQUARES PROBLEMS

Sections in order: 29.1-29.3, 29.5-29.8, 30.1-30.6

MOTIVATION

The earlier parts of this book dealt extensively with the problem of linear *least-mean-squares* estimation, whereby one random variable is estimated from observations of another correlated random variable. For example, in Sec. 8.1 we studied the problem of estimating a zero-mean random variable $\boldsymbol{d}$ from a zero-mean random row vector $\boldsymbol{u}$, by seeking the optimal column vector w that solves

$$\min_{w} \ \mathsf{E}\,|\boldsymbol{d} - \boldsymbol{u}w|^2 \tag{29.1}$$

The optimal estimator was found to be

$$\hat{\boldsymbol{d}} = \boldsymbol{u}w^o \qquad \text{where} \qquad w^o = R_u^{-1} R_{du} \tag{29.2}$$

in terms of the second-order moments of $\{\boldsymbol{d}, \boldsymbol{u}\}$, namely,

$$R_u = \mathsf{E}\,\boldsymbol{u}^*\boldsymbol{u} > 0 \quad \text{and} \quad R_{du} = \mathsf{E}\,\boldsymbol{d}\boldsymbol{u}^*$$

MOTIVATION

The resulting minimum mean-square error was further seen to be given by

$$\text{m.m.s.e.} = \sigma_d^2 - R_{ud}R_u^{-1}R_{du} \;=\; \sigma_d^2 - \sigma_{\hat{d}}^2$$

where $\sigma_d^2 = \mathsf{E}\,|\boldsymbol{d}|^2$ and $\sigma_{\hat{d}}^2 = \mathsf{E}\,|\hat{\boldsymbol{d}}|^2$.

We proceeded in Chapter 8 to devise steepest-descent schemes for evaluating w^o iteratively, and in Chapter 10 we showed how stochastic gradient algorithms can be used to approximate w^o, also iteratively. These latter algorithms were aimed at situations where access to the moments $\{R_{du}, R_u\}$ was not readily available, but only access to realizations $\{d(i), u_i\}$ of the random variables $\{\boldsymbol{d}, \boldsymbol{u}\}$. One such stochastic gradient method was the LMS algorithm (cf. Sec. 10.2):

$$w_i = w_{i-1} \;+\; \mu u_i^*[d(i) - u_i w_{i-1}], \qquad w_{-1} = 0$$

MOTIVATION

Another stochastic gradient method was the exponentially-weighted RLS algorithm described in Sec. 14.1:

$$\begin{aligned} P_i &= \lambda^{-1}\left[P_{i-1} - \frac{\lambda^{-1}P_{i-1}u_i^* u_i P_{i-1}}{1+\lambda^{-1}u_i P_{i-1}u_i^*}\right], \qquad P_{-1} = \epsilon^{-1}\mathrm{I} \\ w_i &= w_{i-1} + P_i\, u_i^*[d(i) - u_i w_{i-1}], \qquad w_{-1} = 0 \end{aligned}$$

for some $0 \ll \lambda \leq 1$. Both LMS and RLS were motivated and derived in Chapter 10 by appealing to instantaneous-gradient approximations.

The purpose of Chapters 29–43 is to study the recursive least-squares algorithm in greater detail. Rather than motivate it as a stochastic gradient *approximation* to a steepest-descent method, as was done in Sec. 14.1, the discussion in these chapters will bring forth deeper insights into the nature of the RLS algorithm. In particular, it will be seen in Chapter 30 that RLS is an *optimal* (as opposed to *approximate*) solution to a well-defined optimization problem. In addition, the discussion will reveal that RLS is very rich in structure, so much so that many equivalent variants exist. While all these variants are mathematically equivalent, they vary among themselves in computational complexity, performance under finite-precision conditions, and even in modularity and ease of implementation.

LEAST-SQUARES PROBLEM

29.1 LEAST-SQUARES PROBLEM

Assume we have available N realizations of the random variables $\boldsymbol{d}$ and $\boldsymbol{u}$, say

$$\{d(0), d(1), \ldots, d(N-1)\} \quad \text{and} \quad \{u_0, u_1, \ldots, u_{N-1}\}$$

respectively, where the $\{d(i)\}$ are scalars and the $\{u_i\}$ are $1 \times M$. Given the $\{d(i), u_i\}$, and assuming ergodicity,[10] we can approximate the mean-square-error cost in (29.1) by its sample average as

$$\mathsf{E}\,|\mathrm{d} - \boldsymbol{u}w|^2 \;\approx\; \frac{1}{N}\sum_{i=0}^{N-1} |d(i) - u_i w|^2 \tag{29.3}$$

In this way, the optimization problem (29.1) can be replaced by the related problem:

$$\min_{w} \left(\sum_{i=0}^{N-1} |d(i) - u_i w|^2 \right) \tag{29.4}$$

where we have removed the scaling factor $1/N$.

[10] As explained before in Chapter 19, an ergodic random process is one for which time-averages coincide with ensemble averages.

LEAST-SQUARES PROBLEM

Vector Formulation

The cost function (29.4) can be reformulated in vector notation as follows. We collect the observations $\{d(i)\}$ into an $N \times 1$ vector y and the row vectors $\{u_i\}$ into an $N \times M$ data matrix H:

$$y \triangleq \begin{bmatrix} d(0) \\ d(1) \\ d(2) \\ \vdots \\ d(N-1) \end{bmatrix}, \qquad H \triangleq \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix}$$

Then (29.4) can be rewritten as

$$\min_{w} \ \|y - Hw\|^2 \tag{29.5}$$

where the notation $\|\cdot\|^2$ denotes the squared Euclidean norm of its argument, namely, $\|a\|^2 = a^*a$ for any column vector a. Problem (29.5) is known as the standard least-squares problem.

LEAST-SQUARES PROBLEM

Definition 29.1 (Least-squares problem) Given an $N \times 1$ vector y and an $N \times M$ data matrix H, the least-squares problem seeks an $M \times 1$ vector w that solves $\min_w \|y - Hw\|^2$.

Two cases can occur depending on the relation between the dimensions $\{N, M\}$:

1. **Over-determined least-squares** ($N \geq M$): In this case, the data matrix H has at least as many rows as columns, so that the number of measurements (i.e., the number of entries in y) is at least equal to the number of unknowns (i.e., the number of entries in w). This situation corresponds to an *over-determined* least-squares problem and, as we shall see, (29.5) will either have a unique solution or an infinite number of solutions.

2. **Under-determined least-squares** ($N < M$): In this case, the data matrix H has fewer rows than columns, so that the number of measurements is less than the number of unknowns. This situation corresponds to an *under-determined* least-squares problem for which (29.5) will have an infinite number of solutions.

LEAST-SQUARES PROBLEM

The purpose of the discussion that follows is to show that all solutions $\widehat{w}$ to the least-squares problem (29.5) are characterized as solutions to the linear system of equations

$$\boxed{H^*H\widehat{w} = H^*y}$$

which are known as the *normal equations*. In addition, the discussion will clarify under what conditions a unique $\widehat{w}$ exists, as opposed to infinitely many, and it will highlight an important orthogonality property of least-squares solutions. In our presentation, we shall use both geometric and algebraic derivations to establish these facts. We start with the geometric argument and later show how to arrive at the same conclusions by means of algebraic arguments.

DIFFERENTIATION ARGUMENT

29.3 ALGEBRAIC ARGUMENTS

Before summarizing the above discussion, and before enumerating additional properties of the least-squares solution, we proceed to re-derive the above results by means of two independent algebraic arguments.

Differentiation Argument

Let $J(w)$ denote the cost function in (29.5), i.e.,

$$J(w) \triangleq \|y - Hw\|^2 = \|y\|^2 - y^*Hw - w^*H^*y + w^*H^*Hw \quad (29.11)$$

Differentiating $J(w)$ with respect to w we find that its gradient vector evaluates to zero at all $\widehat{w}$ that satisfy

$$-y^*H + \widehat{w}^*H^*H = 0$$

which are again the normal equations (29.7). The solution(s) $\widehat{w}$ so obtained correspond to minima of $J(w)$ since its Hessian matrix is nonnegative-definite, i.e.,

$$\nabla_w^2[J(w)] = H^*H \geq 0$$

SOLUTION

29.2 GEOMETRIC ARGUMENT

Our objective is to characterize all solutions of (29.5). We thus note first that, for any w, the vector Hw lies in the column span (or range space) of the data matrix H, written as $Hw \in \mathcal{R}(H)$. Therefore, the least-squares criterion (29.5) is such that it seeks a column vector in the range space of H that is closest to y in the Euclidean norm sense. Specifically, the least-squares problem seeks a $\widehat{w}$ such that $H\widehat{w}$ is closest to y.

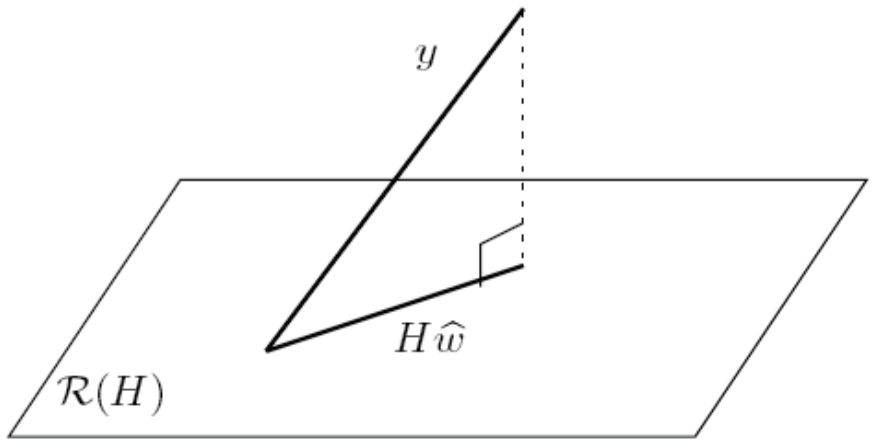

FIGURE 29.1 A least-squares solution is obtained when $y - H\widehat{w}$ is orthogonal to $\mathcal{R}(H)$.

GEOMETRY

Now we know from Euclidean geometry that the closest vector to y within $\mathcal{R}(H)$ is such that the residual vector, $y - H\widehat{w}$, is orthogonal to all vectors in $\mathcal{R}(H)$ as illustrated in Fig. 29.1. For readers not familiar with the geometry of vectors in Euclidean space, the algebraic derivations in Sec. 29.3 will arrive at the same conclusion and will therefore provide a justification for this claim.

Therefore, it must hold that any candidate solution $\widehat{w}$ should result in a residual $(y - H\widehat{w})$ that is orthogonal to Hp, for any vector p or, equivalently, $p^*H^*(y - H\widehat{w}) = 0$. Clearly, the only vector that is orthogonal to any vector p is the zero vector, so that we must have

$$H^*(y - H\widehat{w}) = 0 \tag{29.6}$$

and we conclude that any solution $\widehat{w}$ of the least-squares problem (29.5) must satisfy the so-called normal equations:

$$\boxed{H^*H\widehat{w} = H^*y} \tag{29.7}$$

These equations are always consistent, i.e., a solution $\widehat{w}$ always exists. This is because, as was shown earlier in App. B.2, the matrices H^*H and H^* have the same column span, i.e., $\mathcal{R}(H^*) = \mathcal{R}(H^*H)$.

GEOMETRY

For any solution $\widehat{w}$ of (29.7), we denote the resulting closest vector to y by $\widehat{y} = H\widehat{w}$ and we refer to it as the *projection* of y onto $\mathcal{R}(H)$:

$$\boxed{\widehat{y} = H\widehat{w} \;\triangleq\; \text{projection of } y \text{ onto } \mathcal{R}(H)} \tag{29.8}$$

We show in Thm. 29.2 further ahead that even when the normal equations (29.7) have a multitude of solutions $\widehat{w}$, all of them will lead to the *same* value for $\widehat{y} = H\widehat{w}$. After all, from a geometric point of view, projecting y onto $\mathcal{R}(H)$ results in a unique projection $\widehat{y}$. What the many $\widehat{w}'s$ amount to, when they exist, are equivalent representations for this unique $\widehat{y}$ in terms of the columns of H.

We shall denote the corresponding residual vector by

$$\boxed{\widetilde{y} \;\triangleq\; y - H\widehat{w}}$$

so that the orthogonality condition (29.6) can be rewritten as

$$\boxed{H^*\widetilde{y} = 0} \qquad \text{(orthogonality condition)} \tag{29.9}$$

GEOMETRY

We shall often express this orthogonality condition more succintly as $\widetilde{y} \perp \mathcal{R}(H)$, where the $\perp$ notation is used to mean that $\widetilde{y}$ is orthogonal to any vector in the range space (column span) of H. In particular, since, by construction, $\widehat{y} \in \mathcal{R}(H)$, it also holds that

$$\widetilde{y} \perp \widehat{y} \quad \text{or} \quad \widehat{y}^* \widetilde{y} = 0$$

Let ξ denote the minimum cost of (29.5). It can be evaluated as follows:

$$\begin{aligned}
\xi &= \|y - H\widehat{w}\|^2 \\
&= (y - H\widehat{w})^*(y - H\widehat{w}) \\
&= y^*(y - H\widehat{w}), \quad \text{since } \widehat{w}^* H^* \widetilde{y} = 0 \text{ by (29.9)} \\
&= y^* y - y^* H\widehat{w} \\
&= y^* y - \widehat{w}^* H^* H\widehat{w}, \quad \text{since } y^* H = \widehat{w}^* H^* H \text{ by (29.7)} \\
&= y^* y - \widehat{y}^* \widehat{y}
\end{aligned}$$

That is, we obtain the following equivalent representations for the minimum cost:

$$\boxed{\xi = \|y\|^2 - \|\widehat{y}\|^2 = y^* \widetilde{y}} \tag{29.10}$$

NORMAL EQUATIONS

Theorem 29.1 (The normal equations) A vector $\widehat{w}$ solves the least-squares problem (29.5) if, and only if, it satisfies the normal equations

$$H^* H \widehat{w} = H^* y$$

or, equivalently, if and only if, it satisfies the orthogonality condition

$$y - H\widehat{w} \perp \mathcal{R}(H)$$

The normal equations are always consistent, i.e., a solution $\widehat{w}$ always exists and the resulting minimum cost is given by either expression:

$$\xi = \|y\|^2 - \|\widehat{y}\|^2 = y^* \widetilde{y}$$

where $\widehat{y} = H\widehat{w}$ is the projection of y onto $\mathcal{R}(H)$ and $\widetilde{y} = y - \widehat{y}$ is the residual vector.

PROPERTIES

Theorem 29.2 (Properties of solutions) The solution of the least-squares problem (29.5) has the following properties:

1. The solution $\widehat{w}$ is unique if, and only if, the data matrix H has full column rank (i.e., all its columns are linearly independent, which necessarily requires $N \geq M$). In this case, $\widehat{w}$ is given by

$$\widehat{w} = (H^* H)^{-1} H^* y$$

   This situation occurs only for over-determined least-squares problems.

2. When $H^* H$ is singular, then infinitely many solutions $\widehat{w}$ exist and any two solutions differ by a vector in the nullspace of H, i.e., if $\widehat{w}_1$ and $\widehat{w}_2$ are any two solutions, then $\widehat{w}_1 - \widehat{w}_2 \in \mathcal{N}(H)$. This situation can occur for both over- and under-determined least-squares problems.[13]

3. When many solutions $\widehat{w}$ exist, regardless of which one we pick, the resulting projection vector $\widehat{y} = H\widehat{w}$ is the same and the resulting minimum cost is also the same and given by $\xi = \|y\|^2 - \|\widehat{y}\|^2$.

4. When many solutions $\widehat{w}$ exist, the one that has the smallest Euclidean norm, namely, the one that solves

$$\min_{\widehat{w}} \ \|\widehat{w}\|^2 \quad \text{subject to} \quad H^* H \widehat{w} = H^* y$$

   is given by $\widehat{w} = H^\dagger y$, where $H^\dagger$ denotes the pseudo-inverse of H.[14]

PROOF

We proceed in steps.

1. The normal equations $H^*H\widehat{w} = H^*y$ have a unique solution if, and only if, H^*H is invertible. This condition cannot happen when $N < M$ since the product H^*H will be rank deficient. Therefore, we must have $N \geq M$. Moreover, recall that we proved in App. B.2 that for any matrix H having at least as many rows as columns, it holds that H^*H is invertible if, and only if, H has full rank. These facts establish the first statement in the theorem.

2. Let r be any nonzero vector in the nullspace of H^*H, i.e., $H^*Hr = 0$. If $\widehat{w}$ solves the normal equations, i.e., if $H^*H\widehat{w} = H^*y$, then so does $\widehat{w}+r$ since $H^*H(\widehat{w}+r) = H^*y$. Therefore, infinitely many solutions to the normal equations exist in this case. Moreover, if $\widehat{w}_1$ and $\widehat{w}_2$ are any two solutions, say $H^*H\widehat{w}_1 = H^*y$ and $H^*H\widehat{w}_2 = H^*y$ then $H^*H(\widehat{w}_1 - \widehat{w}_2) = 0$. That is, $\widehat{w}_1 - \widehat{w}_2 \in \mathcal{N}(H^*H)$. However, we proved in App. B.2 that, for any matrix H, the matrices H^*H and H have the same nullspace and, hence, $H(\widehat{w}_1 - \widehat{w}_2) = 0$. These facts establish the second statement in the theorem.

3. Let $\widehat{w}_1$ and $\widehat{w}_2$ denote any two solutions when multiple solutions exist and let $\widehat{y}_1 = H\widehat{w}_1$ and $\widehat{y}_2 = H\widehat{w}_2$ denote the corresponding projections. Then $\widehat{y}_1 - \widehat{y}_2 = H(\widehat{w}_1 - \widehat{w}_2) = 0$ since, by the second property, $\widehat{w}_1 - \widehat{w}_2 \in \mathcal{N}(H)$. Therefore, $\widehat{y}_1 = \widehat{y}_2$, which establishes the third statement in the theorem.

PROOF

4. The fourth statement in the theorem is proven in App. B.6 (see Lemma B.7) in the general case. Here we remark that when H has full rank, its pseudo-inverse is given by the following expressions:

$$H^{\dagger} = \begin{cases} (H^*H)^{-1}H^* & \text{when } N > M \text{ (a "tall" matrix)} \\ H^*(HH^*)^{-1} & \text{when } N < M \text{ (a "fat" matrix)} \\ H^{-1} & \text{when } N = M \text{ (a square matrix)} \end{cases}$$

When H is rank-deficient, it is more convenient to define its pseudo-inverse in terms of its singular value decomposition, as explained in App. B.6. [See also Prob. VII.6 for a proof, from first principles, of the fourth statement of the theorem in the under-determined case.]

$\diamond$

29.5 PROJECTION MATRICES

We restrict ourselves in this section to the case of over-determined least-squares problems with a full-rank data matrix H (and, hence, $N \geq M$). In this case, the coefficient matrix H^*H is invertible (actually positive-definite) and the least-squares problem (29.5) will have a unique solution that is given by

$$\widehat{w} = (H^*H)^{-1}H^*y$$

with the corresponding projection vector

$$\widehat{y} = H\widehat{w} = H(H^*H)^{-1}H^*y$$

The matrix multiplying y in the above expression is called the *projection* matrix and we denote it by

$$\boxed{\mathcal{P}_H \triangleq H(H^*H)^{-1}H^*, \quad \text{when } H \text{ has full column rank}} \tag{29.19}$$

PROJECTION

The designation *projection matrix* stems from the fact that multiplying y by $\mathcal{P}_H$ amounts to projecting it onto the column span of H. Such projection matrices play a prominent role in least-squares theory and they have many useful properties. For example, projection matrices are Hermitian and also idempotent, i.e., they satisfy

$$\boxed{\mathcal{P}_H^* = \mathcal{P}_H, \qquad \mathcal{P}_H^2 = \mathcal{P}_H} \tag{29.20}$$

Note further that the residual vector, $\widetilde{y} = y - H\widehat{w}$, is given by

$$\widetilde{y} \;=\; y - \mathcal{P}_H y \;=\; (\mathrm{I} - \mathcal{P}_H)y \;=\; \mathcal{P}_H^{\perp} y$$

so that the matrix

$$\boxed{\mathcal{P}_H^{\perp} \;\triangleq\; \mathrm{I} - \mathcal{P}_H}$$

is called the projection matrix onto the orthogonal complement space of H. It is also easy

PROJECTION

It is also easy to see that the minimum cost of the least-squares problem (29.5) can be expressed in terms of $\mathcal{P}_H^{\perp}$ as follows:

$$\begin{aligned}
\xi &= y^*y - \widehat{y}^*\widehat{y} \\
&= y^*y - y^*\mathcal{P}_H^*\mathcal{P}_H y \\
&= y^*y - y^*\mathcal{P}_H y, \qquad \text{since } \mathcal{P}_H^*\mathcal{P}_H = \mathcal{P}_H^2 = \mathcal{P}_H \\
&= y^*\mathcal{P}_H^{\perp} y
\end{aligned}$$

Lemma 29.1 (Unique solution) When the matrix H has full-column rank (and, hence, $N \geq M$), the least-squares problem (29.5) will have a unique solution that is given by $\widehat{w} = (H^*H)^{-1}H^*y$. Moreover, the projection of y onto $\mathcal{R}(H)$, and the corresponding residual vector, are given by $\widehat{y} = \mathcal{P}_H y$ and $\widetilde{y} = \mathcal{P}_H^{\perp} y$ so that y can be decomposed as

$$y = \widehat{y} + \widetilde{y} = \mathcal{P}_H y + \mathcal{P}_H^{\perp} y$$

with $\|y\|^2 = \|\widehat{y}\|^2 + \|\widetilde{y}\|^2$. The resulting minimum cost is $\xi = y^*\mathcal{P}_H^{\perp} y$.

29.6 WEIGHTED LEAST-SQUARES

It is often the case that weighting is incorporated into the cost function of the least-squares problem, so that (29.5) is replaced by

$$\min_{w} (y - Hw)^* W (y - Hw) \qquad W > 0 \tag{29.21}$$

where W is a Hermitian positive-definite matrix. For example, when W is diagonal, its elements assign different weights to the entries of the error vector $y - Hw$.

We shall often rewrite the cost function in (29.21) more compactly as

$$\min_{w} \ \|y - Hw\|_W^2 \tag{29.22}$$

where, for any column vector x, the notation $\|x\|_W^2$ refers to the weighted Euclidean norm of x, i.e., $\|x\|_W^2 = x^* W x$.

WEIGHTED LEAST-SQUARES

One way to solve (29.22) is to show that it reduces to the standard form (29.5). To see this, we introduce the eigen-decomposition $W = V\Delta V^*$, where Δ is diagonal with positive entries and V is unitary, i.e., it satisfies $VV^* = V^*V = \mathrm{I}$. Let $\Delta^{1/2}$ denote the diagonal matrix whose entries are equal to the positive square-roots of the entries of Δ, and define the change of variables:

$$a \triangleq \Delta^{1/2}V^*y, \qquad A \triangleq \Delta^{1/2}V^*H \tag{29.23}$$

Observe that since both $\Delta^{1/2}$ and V are invertible, it follows that A has full column rank if, and only if, H has full column rank.

Using the variables $\{a, A\}$ so defined, we can rewrite the weighted problem (29.21) in the equivalent form

$$\min_{w} \ \|a - Aw\|^2 \tag{29.24}$$

which is of the same form as the standard (unweighted) least-squares problem (29.5).

WEIGHTED LEAST-SQUARES

Therefore, we can extend all the results we obtained for (29.5) to the weighted version (29.21) by working with (29.24) instead. In particular, we readily conclude from (29.24) that any solution $\widehat{w}$ to (29.21) should satisfy the orthogonality condition $A^*(a - A\widehat{w}) = 0$ (cf. (29.6)), which, upon using the definitions (29.23) for $\{a, A\}$, can be rewritten in terms of the original data $\{y, H\}$ as

$$H^*V\Delta^{1/2}\left(\Delta^{1/2}V^*y \;-\; \Delta^{1/2}V^*H\widehat{w}\right) = 0$$

or, equivalently,

$$\boxed{H^*W(y - H\widehat{w}) = 0} \tag{29.25}$$

Comparing with the orthogonality condition (29.6) in the unweighted case, we see that the only difference is the presence of the weighting matrix W. This conclusion suggests that we can extend to the weighted least-squares setting the same geometric properties of the standard least-squares setting if we simply employ the concept of *weighted* inner products.

WEIGHTED LEAST-SQUARES

Specifically, for any two column vectors $\{c, d\}$, we can define their weighted inner product as $\langle c, d\rangle_W = c^* W d$, and then say that c and d are orthogonal whenever their weighted inner product is zero. Using this definition, we can interpret (29.25) to mean that the residual vector, $y - H\widehat{w}$, is orthogonal to the column span of H in a weighted sense, i.e.,

$$\langle q, y - H\widehat{w}\rangle_W = 0 \quad \text{for any } q \in \mathcal{R}(H)$$

We further conclude from (29.25) that the normal equations (29.7) are now replaced by

$$\boxed{H^* W H \widehat{w} = H^* W y} \tag{29.26}$$

Proceeding in this manner, and applying the results of the previous sections (especially Thms. 29.1 and 29.2) to (29.24), we arrive at the following statement — see Prob. VII.9.

WEIGHTED LEAST-SQUARES

Theorem 29.3 (Weighted least-squares) A vector $\widehat{w}$ is a solution of the weighted least-squares problem (29.21) if, and only if, it satisfies the normal equations $H^*WH\widehat{w} = H^*Wy$. Moreover, the following properties hold:

1. These normal equations are always consistent, i.e., a solution $\widehat{w}$ always exists.

2. The solution $\widehat{w}$ is unique if, and only if, the data matrix H has full column rank, which necessarily requires $N \geq M$. In this case, $\widehat{w}$ is given by $\widehat{w} = (H^*WH)^{-1}H^*Wy$.

3. When H^*WH is singular, which is equivalent to H^*H being singular, then many solutions $\widehat{w}$ exist and any two solutions differ by a vector in the nullspace of H, i.e., if $\widehat{w}_1$ and $\widehat{w}_2$ are any two solutions, then $H(\widehat{w}_1 - \widehat{w}_2) = 0$.

4. When many solutions $\widehat{w}$ exist, regardless of which one we pick, the resulting projection vector $\widehat{y} = H\widehat{w}$ is the same and the resulting minimum cost is also the same and given by either expression: $\xi = \|y\|_W^2 - \|\widehat{y}\|_W^2 = y^*W\widetilde{y}$, where $\widehat{y} = H\widehat{w}$.

5. When many solutions $\widehat{w}$ exist, the one that has the smallest Euclidean norm, namely, the one that solves

$$\min_{\widehat{w}} \|\widehat{w}\|^2 \quad \text{subject to} \quad H^*WH\widehat{w} = H^*Wy$$

is given by $\widehat{w} = A^{\dagger}a$, where $A = \Delta^{1/2}V^*H$ and $a = \Delta^{1/2}V^*y$.

PROJECTIONS

Note that in the special case of over-determined least-squares problems with full-rank data matrices H (and, hence, $N \geq M$), problem (29.21) will have a unique solution that is given by

$$\widehat{w} = (H^*WH)^{-1}H^*Wy$$

with the corresponding projection vector

$$\widehat{y} = H\widehat{w} = H(H^*WH)^{-1}H^*Wy$$

We shall continue to refer to the matrix multiplying y in the above expression as a *projection* matrix,

$$\boxed{\mathcal{P}_H \triangleq H(H^*WH)^{-1}H^*W, \quad \text{when } H \text{ has full column rank}} \tag{29.27}$$

and write $\widehat{y} = \mathcal{P}_H y$.

PROJECTIONS

In contrast to the unweighted case (29.20), the matrix $\mathcal{P}_H$ now satisfies the properties:

$$\boxed{\mathcal{P}_H^* W = W\mathcal{P}_H, \qquad \mathcal{P}_H^2 = \mathcal{P}_H, \quad \mathcal{P}_H^* W\mathcal{P}_H = W\mathcal{P}_H}$$

and the minimum cost of (29.21) can be expressed in terms of $\mathcal{P}_H^{\perp}$ as follows:

$$\xi \;=\; y^* W y - \widehat{y}^* W \widehat{y} \;=\; y^* W y - y^* W \mathcal{P}_H y \;=\; y^* W \mathcal{P}_H^{\perp} y$$

where $\mathcal{P}_H^{\perp} = \mathrm{I} - \mathcal{P}_H$.

29.7 REGULARIZED LEAST-SQUARES

A second variation of the standard least-squares problem (29.5) is regularized least-squares. In this formulation, we seek a vector $\widehat{w}$ that solves

$$\min_{w} \left[(w-\bar{w})^{*}\Pi(w-\bar{w}) \;+\; \|y-Hw\|^{2} \right] \tag{29.28}$$

where, compared with (29.5), we are now incorporating the so-called regularization term $\|w-\bar{w}\|_{\Pi}^{2}$. Here, Π is a positive-definite matrix, usually a multiple of the identity, and $\bar{w}$ is a given column vector, usually $\bar{w}=0$.

One motivation for using regularization is that it allows us to incorporate some *a priori* information about the solution into the problem statement. Assume, for instance, that we set $\Pi=\delta I$ and choose δ as a large positive number. Then, the first term in the cost function (29.28) becomes dominant and it is not hard to imagine that the cost will be minimized by a vector $\widehat{w}$ that is close to $\bar{w}$ in order to offset the dominant effect of this first term. For this reason, we say that a "large" Π reflects high confidence that $\bar{w}$ is a good guess for the solution $\widehat{w}$. On the other hand, a "small" Π indicates a high degree of uncertainty in the initial guess $\bar{w}$.

REGULARIZATION

Another reason for regularization is that it relieves problems associated with rank deficiency in the data matrix H. To clarify this issue, we first need to solve (29.28). The solution can be obtained in many ways (including, e.g., plain differentiation of the cost function with respect to w as was done in Sec. 29.3). We choose instead to solve (29.28) by showing again how it can be reduced to the solution of a standard least-squares problem of the form (29.5). This line of argument helps clarify the role of the orthogonality condition in regularized least-squares.

REGULARIZATION

Thus introduce the change of variables $z = w - \bar{w}$ and $b = y - H\bar{w}$, so that the regularized cost function in (29.28) becomes

$$\min_z \left[z^* \Pi z + \|b - Hz\|^2 \right] \tag{29.29}$$

Introduce further the eigen-decomposition of Π, say $\Pi = U\Lambda U^*$ where U is unitary and Λ has positive diagonal entries. Let $\Lambda^{1/2}$ denote the diagonal matrix whose entries are the positive square-roots of the entries of Λ. Then we can rewrite the cost in (29.29) in the equivalent form

$$\min_z \left\| \begin{bmatrix} 0 \\ b \end{bmatrix} - \begin{bmatrix} \Lambda^{1/2} U^* \\ H \end{bmatrix} z \right\|^2 \tag{29.30}$$

This problem is now of the same form as the standard least-squares problem (29.5), where the roles of the vector y and the data matrix H are played by

$$\begin{bmatrix} 0 \\ b \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \Lambda^{1/2} U^* \\ H \end{bmatrix}$$

respectively.

REGULARIZATION

The orthogonality condition (29.6) of least-squares solutions then shows that any solution $\widehat{z}$ must satisfy

$$\begin{bmatrix} \Lambda^{1/2}U^* \\ H \end{bmatrix}^* \left(\begin{bmatrix} 0 \\ b \end{bmatrix} - \begin{bmatrix} \Lambda^{1/2}U^* \\ H \end{bmatrix} \widehat{z} \right) = 0 \tag{29.31}$$

which, upon using $\widehat{z} = \widehat{w} - \bar{w}$, leads to the linear system of equations:

$$\boxed{[\Pi + H^*H]\,(\widehat{w} - \bar{w}) = H^*(y - H\bar{w})} \tag{29.32}$$

These equations characterize the solution of the regularized least-squares problem (29.28). When $\bar{w} = 0$, the equations reduce to

$$[\Pi + H^*H]\,\widehat{w} = H^*y \tag{29.33}$$

Compared with the normal equations (29.7) for the standard least-squares problem (29.5), we see that the presence of the positive-definite matrix Π in (29.33) guarantees an invertible coefficient matrix since $\Pi + H^*H > 0$, regardless of whether H is rank-deficient or not and regardless of how the row and column dimensions of H compare to each other.

REGULARIZATION

Using the expression given in Thm. 29.1 for the minimum cost of a standard least-squares problem, we can similarly evaluate the minimum cost of the regularized least-squares problem (29.30) as

$$\begin{aligned}
\xi &= \begin{bmatrix} 0 \\ b \end{bmatrix}^* \left(\begin{bmatrix} 0 \\ b \end{bmatrix} - \begin{bmatrix} \Lambda^{1/2} U^* \\ H \end{bmatrix} \widehat{z} \right) \\
&= b^*(b - H\widehat{z}) \\
&= (y - H\bar{w})^*[(y - H\bar{w}) - H(\widehat{w} - \bar{w})] \\
&= (y - H\bar{w})^*(y - H\widehat{w}) \\
&= (y - H\bar{w})^*\widetilde{y} \qquad (29.34)
\end{aligned}$$

where $\widetilde{y} = y - \widehat{y} = y - H\widehat{w}$.

REGULARIZATION

where $\widetilde{y} = y - \widehat{y} = y - H\widehat{w}$. An equivalent expression for ξ can be obtained as follows:

$$\begin{aligned}
\xi &= (y - H\bar{w})^* \left[(y - H\bar{w}) \; - \; H(\widehat{w} - \bar{w})\right] \\
&= (y - H\bar{w})^* \left[\mathrm{I} - H[\Pi + H^*H]^{-1}H^*\right](y - H\bar{w}) \\
&= (y - H\bar{w})^* \left[\mathrm{I} + H\Pi^{-1}H^*\right]^{-1}(y - H\bar{w}) \qquad (29.35)
\end{aligned}$$

where we used the matrix inversion lemma (5.4) in the last step. Note that expression (29.34) does not include Π explicitly.

REGULARIZATION

Finally, observe that the orthogonality condition (29.31) can be written as

$$\begin{bmatrix} U\Lambda^{1/2} & H^* \end{bmatrix} \begin{bmatrix} -\Lambda^{1/2}U^*\widehat{z} \\ \widetilde{b} \end{bmatrix} = 0$$

where

$$\widetilde{b} = b - H\widehat{z} = (y - H\bar{w}) - H(\widehat{w} - \bar{w}) = y - \widehat{y} = \widetilde{y}$$

Recalling that $\Pi = U\Lambda U^*$ and $\widehat{z} = \widehat{w} - \bar{w}$, the above orthogonality condition can be rewritten more compactly as

$$\boxed{H^*\widetilde{y} = \Pi(\widehat{w} - \bar{w})} \quad \text{(orthogonality condition)} \tag{29.36}$$

In the absence of regularization, i.e., when $\Pi = 0$, the above result reduces to the standard orthogonality condition (29.9), namely, it becomes $H^*\widetilde{y} = 0$.

REGULARIZED LEAST-SQUARES

Theorem 29.4 (Regularized least-squares) The solution of the regularized least-squares problem (29.28) is always unique and given by

$$\widehat{w} = \bar{w} + [\Pi + H^* H]^{-1} H^* (y - H\bar{w})$$

The resulting minimum cost is given by either expression:

$$\xi \;=\; (y - H\bar{w})^* \widetilde{y} \;=\; (y - H\bar{w})^* \left[\mathrm{I} + H\Pi^{-1}H^*\right]^{-1} (y - H\bar{w})$$

where $\widetilde{y} = y - \widehat{y}$ and $\widehat{y} = H\widehat{w}$. Moreover, $\widehat{w}$ satisfies the orthogonality condition $H^*\widetilde{y} = \Pi(\widehat{w} - \bar{w})$.

29.8 WEIGHTED REGULARIZED LEAST-SQUARES

We can combine the formulations of Secs. 29.6 and 29.7 and introduce a weighted regularized least-squares problem. The weighted version of (29.28) would have the form

$$\min_{w} \quad [\,(w-\bar{w})^{*}\Pi(w-\bar{w}) \;+\; (y-Hw)^{*}W(y-Hw)\,] \tag{29.37}$$

where, as before, W is positive-definite. Actually, with $\Pi > 0$, the weighting matrix W can be allowed to be nonnegative-definite. It is easy to verify that all the expressions in Thm. 29.5 further ahead that do not involve an inverse of W will still hold.

Again, the solution of (29.37) can be obtained in many ways (including plain differentiation with respect to w). Here, as before, we choose to solve (29.37) by showing how it reduces to the standard least-squares problem (29.5).

WEIGHTED REGULARIZED LS

it reduces to the standard least-squares problem (29.5). For this purpose, we resort one more time to the eigen-decomposition $W = V\Delta V^*$, and define the normalized quantities $a = \Delta^{1/2}V^*y$ and $A = \Delta^{1/2}V^*H$. Then the weighted regularized problem (29.37) becomes

$$\min_{w} \left[(w-\bar{w})^*\Pi(w-\bar{w}) + \|a - Aw\|^2 \right] \tag{29.38}$$

which is of the same form as the unweighted regularized least-squares problem (29.28). We can therefore invoke Thm. 29.4, and the definitions of $\{a, A\}$ above, to arrive at the following statement, where the orthogonality condition (29.36) is now replaced by

$$\boxed{H^*W\widetilde{y} = \Pi(\widehat{w} - \bar{w})} \quad \text{(orthogonality condition)} \tag{29.39}$$

WEIGHTED REGULARIZED LS

Theorem 29.5 (Weighted regularized least-squares) The solution of the weighted regularized least-squares problem (29.37) is always unique and given by

$$\widehat{w} = \bar{w} + [\Pi + H^* W H]^{-1} H^* W (y - H\bar{w})$$

and the resulting minimum cost is given by

$$\xi = (y - H\bar{w})^* W \widetilde{y} = (y - H\bar{w})^* \left[W^{-1} + H\Pi^{-1}H^*\right]^{-1} (y - H\bar{w})$$

where $\widetilde{y} = y - \widehat{y}$ and $\widehat{y} = H\widehat{w}$. Moreover, $\widehat{w}$ satisfies the orthogonality condition $H^* W \widetilde{y} = \Pi(\widehat{w} - \bar{w})$.

LEAST-SQUARES PROBLEMS

TABLE 29.1 Normal equations associated with several least-squares problems.

Problem	Cost function	Normal equations
Standard least-squares	$\min_w \|y - Hw\|^2$	$H^* H\widehat{w} = H^* y$
Weighted least-squares	$\min_w \|y - Hw\|^2_W,\ W > 0$	$H^* W H\widehat{w} = H^* W y$
Regularized least-squares	$\min_w \|w - \bar{w}\|^2_\Pi + \|y - Hw\|^2$ $\Pi > 0$	$(\Pi + H^* H)(\widehat{w} - \bar{w}) = H^*(y - H\bar{w})$
Weighted regularized least-squares	$\min_w \|w - \bar{w}\|^2_\Pi + \|y - Hw\|^2_W$ $\Pi > 0,\ W \geq 0$	$(\Pi + H^* W H)(\widehat{w} - \bar{w}) = H^* W(y - H\bar{w})$

ORTHOGONALITY CONDITIONS

TABLE 29.2 Orthogonality conditions associated with several least-squares problems. In the statements below, $\widetilde{y} = y - \widehat{y}$ where $\widehat{y} = H\widehat{w}$.

Problem	Cost function	Orthogonality condition
Standard least-squares	$\min_w \|y - Hw\|^2$	$H^*\widetilde{y} = 0$
Weighted least-squares	$\min_w \|y - Hw\|_W^2, \; W > 0$	$H^*W\widetilde{y} = 0$
Regularized least-squares	$\min_w \|w - \bar{w}\|_\Pi^2 + \|y - Hw\|^2$ $\Pi > 0$	$H^*\widetilde{y} = \Pi(\widehat{w} - \bar{w})$
Weighted regularized least-squares	$\min_w \|w - \bar{w}\|_\Pi^2 + \|y - Hw\|_W^2$ $\Pi > 0, \; W \geq 0$	$H^*W\widetilde{y} = \Pi(\widehat{w} - \bar{w})$

MINIMUM COSTS

TABLE 29.3 Minimum costs associated with several least-squares problems. In the statements below, $\widetilde{y} = y - \widehat{y}$ where $\widehat{y} = H\widehat{w}$.

Problem	Cost function	Minimum cost
Standard least-squares	$\min_w \|y - Hw\|^2$	$y^*\widetilde{y}$
Weighted least-squares	$\min_w \|y - Hw\|_W^2,\ W > 0$	$y^*W\widetilde{y}$
Regularized least-squares	$\min_w \|w - \bar{w}\|_\Pi^2 + \|y - Hw\|^2$ $\Pi > 0$	$(y - H\bar{w})^*\widetilde{y}$
Weighted regularized least-squares	$\min_w \|w - \bar{w}\|_\Pi^2 + \|y - Hw\|_W^2$ $\Pi > 0,\ W \geq 0$	$(y - H\bar{w})^*W\widetilde{y}$