50 Least-Squares Problems

We studied in Chapters 29 and 30 the mean-square error (MSE) criterion in some detail, and applied it to the problem of inferring an unknown (or hidden) variable $\boldsymbol{x}$ from the observation of another variable $\boldsymbol{y}$ when $\{\boldsymbol{x}, \boldsymbol{y}\}$ are related by means of a linear regression model or a state-space model. In the latter case, we derived several algorithms for the solution of the inference problem, such as the Kalman filter, its measurement and time-update forms, and its approximate nonlinear forms. We revisit the linear least-mean-square error (LLMSE) formulation in this chapter and use it to motivate an alternative least-squares method that is purely *data-driven*. This second method will not require knowledge of statistical moments of the variables involved because it will operate directly on data measurements to learn the hidden variable. This data-driven approach to inference will be prevalent in all chapters in this volume, where we describe many other learning algorithms for the solution of general inference problems that rely on other choices for the loss function, other than the quadratic loss.

We start our analysis of data-driven methods by focusing on the least-squares problem because it is mathematically tractable and sheds useful insights on many challenges that will hold more generally. We will explain how some of these challenges are addressed in least-squares formulations (e.g., by using regularization) and subsequently apply similar ideas to other inference problems, especially in the classification context when $\boldsymbol{x}$ assumes discrete values.

50.1 MOTIVATION

The MSE problem of estimating a scalar random variable $\boldsymbol{x} \in \mathbb{R}$ from observations of a vector random variable $\boldsymbol{y} \in \mathbb{R}^M$ seeks a mapping $c(\boldsymbol{y})$ that solves

$$\widehat{\boldsymbol{x}} = \underset{c(\boldsymbol{y})}{\text{argmin}} \; \mathbb{E}\,(\boldsymbol{x} - c(\boldsymbol{y}))^2 \tag{50.1}$$

We showed in (27.18) that the optimal estimate is given by the conditional mean $\widehat{x} = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y)$. For example, for continuous random variables, the MSE estimate involves an integral computation of the form:

$$\widehat{x} \;=\; \int_{x\in\mathcal{X}} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \tag{50.2}$$

over the domain of the realizations, $x \in \mathcal{X}$. Evaluation of this solution requires knowledge of the conditional distribution, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. Even if $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ were available, computation of the integral expression is generally not possible in closed form. In Chapter 29, we limited $c(\boldsymbol{y})$ to the class of *affine* functions of $\boldsymbol{y}$ and considered instead the problem:

$$\boxed{\begin{aligned}(w^o,\theta^o) = \underset{w,\theta}{\text{argmin}}\ &\mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})^2\\ \text{subject to}\ \ &\widehat{\boldsymbol{x}} = \boldsymbol{y}^{\mathsf{T}}w-\theta\end{aligned}} \tag{50.3}$$

for some vector parameter $w \in \mathbb{R}^M$ and offset $\theta \in \mathbb{R}$. The minus sign in front of θ is for convenience. Let $\{\bar{x},\bar{y}\}$ denote the first-order moments of the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., their means:

$$\bar{x} = \mathbb{E}\,\boldsymbol{x}, \qquad \bar{y} = \mathbb{E}\,\boldsymbol{y} \tag{50.4a}$$

and let $\{\sigma_x^2, R_y, r_{xy}\}$ denote their second-order moments, i.e., their (co)-variances and cross-covariance:

$$\sigma_x^2 = \mathbb{E}\,(\boldsymbol{x}-\bar{x})^2 \tag{50.4b}$$

$$R_y = \mathbb{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} \tag{50.4c}$$

$$r_{xy} = \mathbb{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} = r_{yx}^{\mathsf{T}} \tag{50.4d}$$

Theorem 29.1 showed that the LLMSE estimator and the resulting minimum mean-square error (MMSE) are given by

$$\widehat{\boldsymbol{x}}_{\text{LLMSE}} - \bar{x} = r_{xy}R_y^{-1}(\boldsymbol{y}-\bar{y}) \tag{50.5a}$$

$$\text{MMSE} = \sigma_x^2 - r_{xy}R_y^{-1}r_{yx} \tag{50.5b}$$

In other words, the optimal parameters are given by

$$w^o = R_y^{-1}r_{yx}, \quad \theta^o = \bar{y}^{\mathsf{T}}w^o - \bar{x} \tag{50.6}$$

Note in particular that the offset parameter is unnecessary if the variables have zero mean since in that case $\theta^o = 0$. More importantly, observe that the estimator $\widehat{\boldsymbol{x}}_{\text{LLMSE}}$ requires knowledge of the first- and second-order moments of the random variables $\{\boldsymbol{x},\boldsymbol{y}\}$. When this information is not available, we need to follow a different route to solve the inference problem. To do so, we will replace the stochastic risk that appears in (50.3) by an *empirical risk* as follows:

$$(w^\star,\theta^\star) = \underset{w,\theta}{\text{argmin}} \left\{ P(w,\theta) \;\stackrel{\Delta}{=}\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - (y_n^{\mathsf{T}}w-\theta)\Big)^2 \right\} \tag{50.7}$$

which is written in terms of a collection of N independent realizations $\{x(n), y_n\}$; these measurements are assumed to arise from the underlying joint distribution for the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ and they are referred to as the *training data* because they will be used to determine the solution $(w^\star, \theta^\star)$. Once $(w^\star, \theta^\star)$ are learned, they can then be used to predict the x-value corresponding to some future observation y by using

$$\widehat{\boldsymbol{x}} = \boldsymbol{y}^{\mathsf{T}} w^\star - \theta^\star \tag{50.8}$$

Obviously, under ergodicity, the empirical risk in (50.7) converges to the stochastic risk in (50.3) as $N \to \infty$. However, even if ergodicity does not hold, we can still pose the empirical risk minimization problem (50.7) independently and seek its solution. Note that we are denoting the empirical risk by the letter $P(\cdot)$; in this case, it depends on two parameters: w and θ. We are also denoting the optimal parameter values by $(w^\star, \theta^\star)$ to distinguish them from (w^o, θ^o). As explained earlier in the text, we use the $\star$ superscript to refer to minimizers of empirical risks, and the o superscript to refer to minimizers of stochastic risks.

50.1.1 Stochastic Optimization

At this stage, one could consider learning the $(w^\star, \theta^\star)$ by applying any of the stochastic optimization algorithms studied in earlier chapters, such as applying a stochastic gradient algorithm or a mini-batch version of it, say,

$$\left\{\begin{array}{l} \text{select a sample } \{\boldsymbol{x}(n), \boldsymbol{y}_n\} \text{ at random at iteration } n \\ \text{let } \widehat{\boldsymbol{x}}(n) = \boldsymbol{y}_n^{\mathsf{T}} \boldsymbol{w}_{n-1} - \boldsymbol{\theta}(n-1) \\ \text{update } \boldsymbol{w}_n = \boldsymbol{w}_{n-1} + 2\mu \boldsymbol{y}_n(\boldsymbol{x}(n) - \widehat{\boldsymbol{x}}(n)) \\ \text{update } \boldsymbol{\theta}(n) = \boldsymbol{\theta}(n-1) - 2\mu(\boldsymbol{x}(n) - \widehat{\boldsymbol{x}}(n)) \end{array}\right. \tag{50.9}$$

This construction is based on using an instantaneous gradient *approximation* at iteration n. The recursions can be grouped together as follows:

$$\widehat{\boldsymbol{x}}(n) = \begin{bmatrix} 1 & \boldsymbol{y}_n^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} -\boldsymbol{\theta}(n-1) \\ \boldsymbol{w}_{n-1} \end{bmatrix} \tag{50.10a}$$

$$\begin{bmatrix} -\boldsymbol{\theta}(n) \\ \boldsymbol{w}_n \end{bmatrix} = \begin{bmatrix} -\boldsymbol{\theta}(n-1) \\ \boldsymbol{w}_{n-1} \end{bmatrix} + 2\mu\Big(\boldsymbol{x}(n) - \widehat{\boldsymbol{x}}(n)\Big) \begin{bmatrix} 1 \\ \boldsymbol{y}_n \end{bmatrix} \tag{50.10b}$$

which are expressed in terms of the extended variables of dimension $M+1$ each:

$$y' \overset{\Delta}{=} \begin{bmatrix} 1 \\ y \end{bmatrix}, \quad w' = \begin{bmatrix} -\theta \\ w \end{bmatrix} \tag{50.11}$$

Using the extended notation we can write down the equivalent representation:

$$\widehat{\boldsymbol{x}}(n) = (\boldsymbol{y}_n')^{\mathsf{T}} \boldsymbol{w}_n' \tag{50.12a}$$

$$\boldsymbol{w}_n' = \boldsymbol{w}_{n-1}' + 2\mu \boldsymbol{y}_n'(\boldsymbol{x}(n) - \widehat{\boldsymbol{x}}(n)) \tag{50.12b}$$

After sufficient iterations, the estimators $(\boldsymbol{w}_n, \boldsymbol{\theta}(n))$ approach $(w^\star, \theta^\star)$. These values can then be used to predict the hidden variable $x(t)$ for any new observation y_t as follows:

$$\widehat{x}(t) = y_t^{\mathsf{T}} w^\star - \theta^\star \tag{50.13}$$

It turns out, however, that problem (50.7) has a special structure that can be exploited to motivate a second *exact* (rather than approximate) recursive solution, for updating $\boldsymbol{w}_{n-1}$ to $\boldsymbol{w}_n$, known as the *recursive least-squares* (RLS) algorithm.

50.1.2 Least-Squares Risk

Using the extended notation, we rewrite the empirical risk problem (50.7) in the form

$$(w')^\star = \underset{w' \in \mathbb{R}^{M+1}}{\text{argmin}} \left\{ P(w') \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - (y'_n)^\mathsf{T} w'\Big)^2 \right\} \tag{50.14}$$

without an offset parameter. For simplicity of notation, we will assume henceforth that the vectors (w, y_n) have been extended according to (50.11) and will continue to use the same notation (w, y_n), without the prime subscript, for the extended quantities:

$$y \leftarrow \begin{bmatrix} 1 \\ y \end{bmatrix}, \quad w \leftarrow \begin{bmatrix} -\theta \\ w \end{bmatrix} \tag{50.15}$$

We will also continue to denote their dimension generically by M (rather than $M+1$). Thus, our problem becomes one of solving

$$w^\star = \underset{w \in \mathbb{R}^{M}}{\text{argmin}} \left\{ P(w) \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^\mathsf{T} w\Big)^2 \right\} \tag{50.16}$$

from knowledge of N data pairs $\{x(n), y_n\}$. We can rewrite this problem in a more familiar least-squares form by collecting the data into convenient vector and matrix quantities. For this purpose, we introduce the $N \times M$ and $N \times 1$ variables

$$H \triangleq \begin{bmatrix} y_0^\mathsf{T} \\ y_1^\mathsf{T} \\ y_2^\mathsf{T} \\ \vdots \\ y_{N-1}^\mathsf{T} \end{bmatrix}, \quad d \triangleq \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{bmatrix} \tag{50.17}$$

The matrix H contains all observation vectors $\{y_n\}$ transposed as rows, while the vector d contains all target signals $\{x(n)\}$. Then, the risk function takes the form

$$P(w) = \frac{1}{N}\|d - Hw\|^2 \tag{50.18}$$

in terms of the squared Euclidean norm of the error vector $d - Hw$. The scaling by $1/N$ does not affect the location of the minimizer $w^\star$ and, therefore, it can be ignored. In this way, formulation (50.16) becomes the standard least-squares problem:

$$\boxed{w^\star \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\text{argmin}} \; \|d - Hw\|^2} \qquad (\textbf{standard least-squares}) \tag{50.19}$$

We motivated (50.19) by linking it to the MSE formulation (50.3) and replacing the stochastic risk by an empirical risk. Of course, the least-squares problem is of independent interest in its own right. Given a collection of data points $\{x(n), y_n\}$, with scalars $x(n)$ and column vectors y_n, we can formulate problem (50.19) directly in terms of these quantities and seek the vector w that matches Hw to d in the least-squares sense.

Example 50.1 **(Maximum-likelihood interpretation)** There is another way to motivate the least-squares problem as the solution to a maximum-likelihood (ML) estimation problem in the presence of Gaussian noise. Assume we collect N iid observations $\{\boldsymbol{x}(n), \boldsymbol{y}_n\}$, for $0 \le n \le N-1$. Assume further that these observations happen to satisfy a linear regression model of the form:

$$\boldsymbol{x}(n) = \boldsymbol{y}_n^\mathsf{T} w + \boldsymbol{v}(n) \tag{50.20}$$

for some unknown vector $w \in \mathbb{R}^M$, and where $\boldsymbol{v}(n)$ is white Gaussian noise with zero mean and variance σ_v^2, i.e., $\boldsymbol{v} \sim \mathcal{N}_{\boldsymbol{v}}(0, \sigma_v^2)$. It is straightforward to conclude that the likelihood function of the joint observations $\{\boldsymbol{x}(n), \boldsymbol{y}_n\}$ given the model w, is

$$\begin{aligned}
& f_{\boldsymbol{x},\boldsymbol{y}}\left(y_0, \ldots, y_{N-1}, x(0), \ldots, x(N-1); w\right) \\
&= f_{\boldsymbol{v}}(v(0), \ldots, v(N-1); w) \\
&= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left\{ -\frac{\left(x(n) - y_n^\mathsf{T} w\right)^2}{2\sigma_v^2} \right\} \\
&= \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma_v^2} \sum_{n=0}^{N-1} \Big(x(n) - y_n^\mathsf{T} w\Big)^2 \right\}
\end{aligned} \tag{50.21}$$

so that the log-likelihood function is given by

$$\ell\left(\{x(n), y_n\}; w\right) = -\frac{N}{2}\ln(2\pi\sigma_v^2) \; - \; \frac{1}{2\sigma_v^2} \sum_{n=0}^{N-1} \Big(x(n) - y_n^\mathsf{T} w\Big)^2 \tag{50.22}$$

The maximization of the log-likelihood function over w leads to the equivalent problem

$$w^\star = \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ \sum_{n=0}^{N-1} \Big(x(n) - y_n^\mathsf{T} w\Big)^2 \right\} \tag{50.23}$$

which is the same least-squares problem (50.16). In Prob. 50.6 we consider a variation of this argument in which the noise process $\boldsymbol{v}(n)$ is not white, which will then lead to the solution of a *weighted* least-squares problem.

Cambridge University Press & Assessment
978-1-009-21828-3 — Inference and Learning from Data
Ali H. Sayed
Excerpt
More Information

50.2 NORMAL EQUATIONS

Problem (50.19) can be solved in closed form using either algebraic or geometric arguments. We expand the least-squares risk:

$$\|d - Hw\|^2 \;=\; \|d\|^2 - 2d^\mathsf{T}Hw + w^\mathsf{T}H^\mathsf{T}Hw \tag{50.24}$$

and differentiate with respect to w to find that the minimizer $w^\star$ should satisfy the normal equations:

$$\boxed{H^\mathsf{T}Hw^\star = H^\mathsf{T}d} \qquad (\textbf{normal equations}) \tag{50.25}$$

Alternatively, we can pursue a geometric argument to arrive at this same conclusion. Note that, for any w, the vector Hw lies in the column span (or range space) of H, written as $Hw \in \mathcal{R}(H)$. Therefore, the least-squares criterion (50.19) is in effect seeking a column vector in the range space of H that is closest to d in the Euclidean norm sense. We know from Euclidean geometry that the closest vector to d within $\mathcal{R}(H)$ can be obtained by projecting d onto $\mathcal{R}(H)$, as illustrated in Fig. 50.1. This means that the residual vector, $d - Hw^\star$, should be orthogonal to all vectors in $\mathcal{R}(H)$:

$$d - Hw^\star \;\perp Hp, \;\; \text{for any } p \tag{50.26}$$

which is equivalent to

$$p^\mathsf{T}H^\mathsf{T}(d - Hw^\star) = 0, \;\; \text{for any } p \tag{50.27}$$

Clearly, the only vector that is orthogonal to any p is the zero vector, so that

$$H^\mathsf{T}(d - Hw^\star) = 0 \tag{50.28}$$

and we arrive again at the normal equations (50.25).

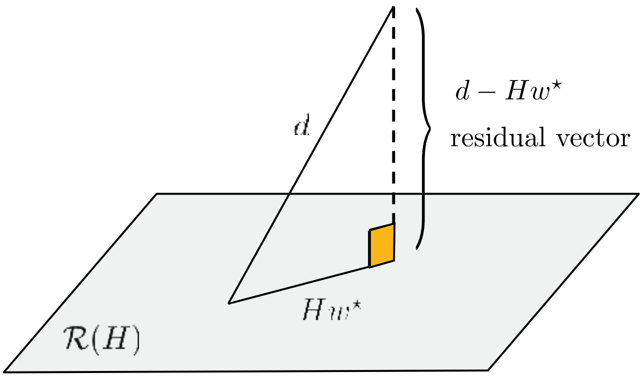

Figure 50.1 A least-squares solution is obtained when $d - Hw^\star$ is orthogonal to $\mathcal{R}(H)$.

50.2.1 Consistent Equations

We explained earlier in Section 1.51 that equations of the form (50.25) are always consistent (i.e., they always have a solution). This is because the matrices H^{T} and $H^{\mathsf{T}}H$ have the same range spaces so that, for any d and H:

$$H^{\mathsf{T}}d \in \mathcal{R}(H^{\mathsf{T}}H) \tag{50.29}$$

Moreover, the normal equations will either have a unique solution or infinitely many solutions. The solution will be unique when $H^{\mathsf{T}}H$ is invertible, which happens when H has full column rank. This condition requires $N \geq M$, which means that there should be at least as many observations as the number of unknowns in w. The full-rank condition implies that the columns of H are not redundant. In this case, we obtain

$$w^{\star} \;=\; (H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}d \tag{50.30}$$

In all other cases, the matrix product $H^{\mathsf{T}}H$ will be rank-deficient. For instance, this situation arises when $N < M$, which corresponds to the case in which we have insufficient data (fewer measurements than the number of unknowns). This situation is not that uncommon in practice. For example, it arises in streaming data implementations when we have not collected enough data to surpass M. When $H^{\mathsf{T}}H$ is singular, the normal equations (50.25) will have infinitely many solutions, all of them differing from each other by vectors in the nullspace of H – recall (1.56). That is, for any two solutions $\{w_1^{\star}, w_2^{\star}\}$ to (50.25), it will hold that

$$w_2^{\star} = w_1^{\star} + p, \quad \text{for some } p \in \mathcal{N}(H) \tag{50.31}$$

Although unnecessary for the remainder of the discussions in this chapter, we explain in Appendix 50.A that when infinitely many solutions $w^{\star}$ exist to the least-squares problem (50.19), we can determine the solution with the smallest Euclidean norm among these by employing the pseudo-inverse of H – see expression (50.179). Specifically, the solution to the following problem

$$\min_{w\in\mathbb{R}^M} \|w\|^2, \quad \text{subject to } H^{\mathsf{T}}Hw = H^{\mathsf{T}}d \tag{50.32}$$

is given by

$$w^{\star} = H^{\dagger}d \tag{50.33}$$

where $H^{\dagger}$ denotes the pseudo-inverse matrix.

50.2.2 Minimum Risk

For any solution $w^{\star}$ of (50.25), we denote the resulting closest vector to d by $\widehat{d} = Hw^{\star}$ and refer to it as the *projection* of d onto $\mathcal{R}(H)$:

$$\widehat{d} = Hw^{\star} \;\stackrel{\Delta}{=}\; \text{projection of } d \text{ onto } \mathcal{R}(H) \tag{50.34}$$

It is straightforward to verify that even when the normal equations have a multitude of solutions, $w^\star$, all of them will lead to the *same* value for $\widehat{d}$. This observation can be justified both algebraically and geometrically. From a geometric point of view, projecting d onto $\mathcal{R}(H)$ results in a unique projection $\widehat{d}$. From an algebraic point of view, if $w_1^\star$ and $w_2^\star$ are two arbitrary solutions, then from (50.31) we find that

$$\widehat{d}_2 \overset{\Delta}{=} Hw_2^\star = H(w_1^\star + p) = Hw_1^\star = \widehat{d}_1 \tag{50.35}$$

What the different solutions $w^\star$ amount to, when they exist, are equivalent representations for the unique $\widehat{d}$ in terms of the columns of H.

We denote the residual vector resulting from the projection by

$$\widetilde{d} \overset{\Delta}{=} d - Hw^\star \tag{50.36}$$

so that the orthogonality condition (50.28) can be rewritten as

$$\boxed{H^\mathsf{T}\widetilde{d} = 0} \qquad (\textbf{orthogonality condition}) \tag{50.37}$$

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

$$\widetilde{d} \perp \widehat{d} \quad \text{or} \quad (\widehat{d})^\mathsf{T}\widetilde{d} = 0 \tag{50.38}$$

Let ξ denote the minimum risk value, i.e., the minimum value of (50.19). This is sometimes referred to as the *training error* because it is the minimum value evaluated on the training data $\{x(n), y_n\}$. It can be evaluated as follows:

$$\begin{aligned}
\xi &= \|d - Hw^\star\|^2 \\
&= (d - Hw^\star)^\mathsf{T}(d - Hw^\star) \\
&= (d - Hw^\star)^\mathsf{T}(d - \widehat{d}) \\
&= d^\mathsf{T}(d - Hw^\star), \quad \text{since } (d - Hw^\star) \perp \widehat{d} \text{ by } (50.38) \\
&= d^\mathsf{T}d - d^\mathsf{T}Hw^\star \\
&= d^\mathsf{T}d - (w^\star)^\mathsf{T}H^\mathsf{T}Hw^\star, \quad \text{since } d^\mathsf{T}H = (w^\star)^\mathsf{T}H^\mathsf{T}H \text{ by } (50.25) \\
&= d^\mathsf{T}d - (\widehat{d})^\mathsf{T}\widehat{d}
\end{aligned} \tag{50.39}$$

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

$$\boxed{\xi = \|d\|^2 - \|\widehat{d}\|^2 = d^\mathsf{T}\widetilde{d}} \qquad (\textbf{minimum risk}) \tag{50.40}$$

50.2.3 Projections

When H has full column rank (and, hence, $N \geq M$), the coefficient matrix $H^\mathsf{T}H$ becomes invertible and the least-squares problem (50.19) will have a unique solution given by

$$w^\star = (H^\mathsf{T} H)^{-1} H^\mathsf{T} d \tag{50.41}$$

with the corresponding projection vector

$$\widehat{d} = Hw^\star = H(H^\mathsf{T} H)^{-1} H^\mathsf{T} d \tag{50.42}$$

The matrix multiplying d in the above expression is called the *projection* matrix onto $\mathcal{R}(H)$ and we denote it by

$$\mathcal{P}_H \;\overset{\Delta}{=}\; H(H^\mathsf{T} H)^{-1} H^\mathsf{T}, \qquad \text{when } H \text{ has full column rank} \tag{50.43}$$

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

$$\mathcal{P}_H^\mathsf{T} = \mathcal{P}_H, \qquad \mathcal{P}_H^2 = \mathcal{P}_H \tag{50.44}$$

Note further that the residual vector, $\widetilde{d} = d - Hw^\star$, is given by

$$\widetilde{d} = d - \mathcal{P}_H d \;=\; (I - \mathcal{P}_H)d \;=\; \mathcal{P}_H^\perp d \tag{50.45}$$

so that the matrix

$$\mathcal{P}_H^\perp \;\overset{\Delta}{=}\; I - \mathcal{P}_H \tag{50.46}$$

is called the projection matrix onto the orthogonal complement space of H. It is easy to see that the minimum risk value can be expressed in terms of $\mathcal{P}_H^\perp$ as follows:

$$\begin{aligned}
\xi &= d^\mathsf{T} d - (\widehat{d})^\mathsf{T} \widehat{d} \\
&= d^\mathsf{T} d - d^\mathsf{T} \mathcal{P}_H^\mathsf{T} \mathcal{P}_H d \\
&= d^\mathsf{T} d - d^\mathsf{T} \mathcal{P}_H d, \qquad \text{since } \mathcal{P}_H^\mathsf{T} \mathcal{P}_H = \mathcal{P}_H^2 = \mathcal{P}_H
\end{aligned} \tag{50.47}$$

That is,

$$\xi \;=\; d^\mathsf{T} \mathcal{P}_H^\perp d \tag{50.48}$$

In summary, we arrive at the following statement for the solution of the standard least-squares problem.

Theorem 50.1. (Solution of least-squares problem) *Consider the standard least-squares problem (50.19) where $H \in \mathbb{R}^{N\times M}$:*

(a) When H has full column rank, which necessitates $N \geq M$, the least-squares problem will have a unique solution given by $w^\star = (H^\mathsf{T}H)^{-1}H^\mathsf{T}d$.

(b) Otherwise, the least-squares problem will have infinitely many solutions $w^\star$ satisfying $H^\mathsf{T}Hw^\star = H^\mathsf{T}d$. Moreover, any two solutions will differ by vectors in $\mathcal{N}(H)$ and the solution with the smallest Euclidean norm is given by $w^\star = H^\dagger d$.

In either case, the projection of d onto $\mathcal{R}(H)$ is unique and given by $\widehat{d} = Hw^\star$. Moreover, the minimum risk value is $\xi = d^\mathsf{T}\widetilde{d}$, where $\widetilde{d} = d - \widehat{d}$.

50.2.4 Weighted and Regularized Variations

There are several extensions and variations of the least-squares formulation, which we will encounter at different locations in our treatment. For example, one may consider a *weighted* least-squares problem of the form

$$w^\star \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{(d-Hw)^\mathsf{T}R(d-Hw)\right\} \qquad (\textbf{weighted least-squares}) \tag{50.49}$$

where $R \in \mathbb{R}^{N\times N}$ is a symmetric positive-definite weighting matrix. Assume, for illustration purposes, that R is diagonal with entries $\{r(n)\}$. Then, the above problem reduces to (we prefer to restore the $1/N$ factor when using the original data):

$$w^\star \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{\frac{1}{N}\sum_{n=0}^{N-1} r(n)\Big(x(n) - y_n^\mathsf{T}w\Big)^2\right\} \tag{50.50}$$

where the individual squared errors appear scaled by $r(n)$. In this way, errors originating from some measurements will be scaled more or less heavily than errors originating from other measurements. In other words, incorporating a weighting matrix R into the least-squares formulation allows the designer to control the relative importance of the errors contributing to the risk value.

One can also consider penalizing the size of the parameter w by modifying the weighted risk function in the following manner:

($\boldsymbol{\ell_2}$**-regularized weighted least-squares**)

$$w^\star \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{\rho\|w\|^2 \;+\; (d-Hw)^\mathsf{T}R(d-Hw)\right\} \tag{50.51}$$

where $\rho > 0$ is called an ℓ_2-regularization parameter (since it penalizes the ℓ_2-norm of w). We will discuss regularization in greater detail in the next chapter. Here, we comment briefly on its role. Observe, for instance, that if ρ is large, then the term $\rho\|w\|^2$ will have a nontrivial effect on the value of the risk function. As such, when ρ is large, the solution $w^\star$ should have smaller Euclidean norm