50 Least-Squares Problems

We studied in Chapters 29 and 30 the mean-square error 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 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 mean-square-error (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 vector:

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

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

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

Theorem 29.1 showed that the linear least-mean-square error (l.l.m.s.e.) estimator and the resulting minimum mean-square error (m.m.s.e.) are given by

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

$$\text{m.m.s.e.} = \sigma_x^2 - r_{xy}R_y^{-1}r_{yx} \tag{50.5b}$$

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 can 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:

$$\widehat{\boldsymbol{x}}(n) = \left[\begin{array}{cc} 1 & \boldsymbol{y}_n^\mathsf{T} \end{array}\right] \left[\begin{array}{c} -\boldsymbol{\theta}(n-1) \\ \boldsymbol{w}_{n-1} \end{array}\right] \tag{50.10a}$$

$$\left[\begin{array}{c} -\boldsymbol{\theta}(n) \\ \boldsymbol{w}_n \end{array}\right] = \left[\begin{array}{c} -\boldsymbol{\theta}(n-1) \\ \boldsymbol{w}_{n-1} \end{array}\right] + 2\mu\Big(\boldsymbol{x}(n) - \widehat{\boldsymbol{x}}(n)\Big) \left[\begin{array}{c} 1 \\ \boldsymbol{y}_n \end{array}\right] \tag{50.10b}$$

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

$$y' \overset{\Delta}{=} \left[\begin{array}{c} 1 \\ y \end{array}\right], \quad w' = \left[\begin{array}{c} -\theta \\ w \end{array}\right] \tag{50.11}$$

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

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

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

$$w^\star = \underset{w}{\text{argmin}} \left\{ P(w) \;\overset{\Delta}{=}\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^\mathsf{T}w\Big)^2 \right\} \tag{50.16}$$

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

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

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 mean-square error 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 estimation problem in the presence of Gaussian noise. Assume we collect N independent and identically-distributed observations $\{\boldsymbol{x}(n), \boldsymbol{y}_n\}$, for $0 \leq n \leq N-1$. 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., $\boldsymbol{v} \sim \mathbb{N}_{\boldsymbol{v}}(0, \sigma_v^2)$. It is straightforward to conclude that the likelihood function of the joint observations $\{\boldsymbol{x}(n), \boldsymbol{y}_n\}$ given the model w is

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

so that the log-likelihood function is given by

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

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

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

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.

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

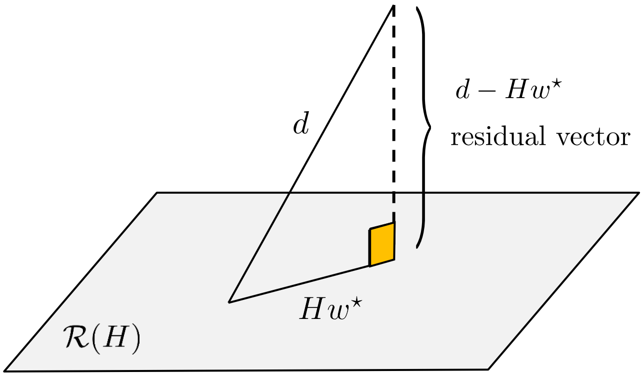

Figure 50.1 A least-squares solution is obtained when $d - Hw^\star$ is orthogonal to $\mathcal{R}(H)$.

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

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).

50.2.1 Consistent Equations

We explained earlier in Sec. 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 (less 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 $\mathcal{R}(H)$:

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

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

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

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

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

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

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

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

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

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 $H^\mathsf{T} H w^\star = H^\mathsf{T} d$. 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

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

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):

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

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:

$$(\boldsymbol{\ell_2}-\textbf{regularized weighted least-squares})$$
$$w^\star \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{\rho\|w\|^2 \;+\; (d - Hw)^\mathsf{T} R(d - Hw)\right\} \tag{50.51}$$

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 since the objective is to minimize the overall risk. In this way, the parameter ρ provides the designer with the flexibility to limit the norm of w to small values. Additionally, it is straightforward to verify by differentiating the above risk function that the solution $w^\star$ satisfies the equations:

$$(\rho I_M + H^\mathsf{T} R H) w^\star = H^\mathsf{T} R d \tag{50.52}$$

Observe, in particular, that even when the product $H^\mathsf{T} R H$ happens to be singular, the coefficient matrix $\rho I_M + H^\mathsf{T} R H$ will be positive-definite and, hence, invertible, due to the addition of the positive term ρI_M. This ensures that the solution will always be unique and given by

$$w^\star = (\rho I_M + H^\mathsf{T} R H)^{-1} H^\mathsf{T} R d \tag{50.53}$$

Example 50.2 (**Sea level change**) We apply the least-squares formalism to the problem of fitting a regression line through measurements related to the change in sea level (measured in mm) relative to the start of year 1993. There are $N = 952$ data points consisting of fractional year values and the corresponding sea level change. We denote the fractional year value by $y(n)$ and the sea level change by $x(n)$ for every entry $n = 1, 2, \ldots, 952$. For example, the second entry ($n = 2$) in the data corresponds to year 1993.0386920, which represents a measurement performed about 14 days into year 1993.

Using the least-squares formalism, we already know how to fit a regression line through these data points by solving a problem of the form:

$$(\alpha^\star, \theta^\star) \triangleq \underset{\alpha,\theta}{\text{argmin}} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} \Big(x(n) - (\alpha y(n) - \theta) \Big)^2 \right\} \tag{50.54}$$

where (α, θ) are scalar parameters in this case. For convenience, we employ the vector notation as follows. We collect the measurements $\{x(n), y(n)\}$ into the $N \times 1$ vector and $N \times 2$ matrix quantities:

$$d = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} \in \mathbb{R}^N, \quad H = \begin{bmatrix} 1 & y(0) \\ 1 & y(1) \\ \vdots & \\ 1 & y(N-1) \end{bmatrix} \in \mathbb{R}^{N\times 2} \tag{50.55}$$

and introduce the parameter vector:

$$w \triangleq \begin{bmatrix} -\theta \\ \alpha \end{bmatrix} \in \mathbb{R}^2 \tag{50.56}$$

Then, problem (50.54) is equivalent to

$$w^\star \triangleq \underset{w\in\mathbb{R}^2}{\text{argmin}} \; \|d - Hw\|^2 \tag{50.57}$$

whose solution is given by

$$w^\star = (H^\mathsf{T} H)^{-1} H^\mathsf{T} d \triangleq \begin{bmatrix} -\theta^\star \\ \alpha^\star \end{bmatrix} \tag{50.58}$$

We find that

$$w^\star \approx \begin{bmatrix} -5961.9 \\ 2.9911 \end{bmatrix} = \begin{bmatrix} -\theta^\star \\ \alpha^\star \end{bmatrix} \tag{50.59}$$

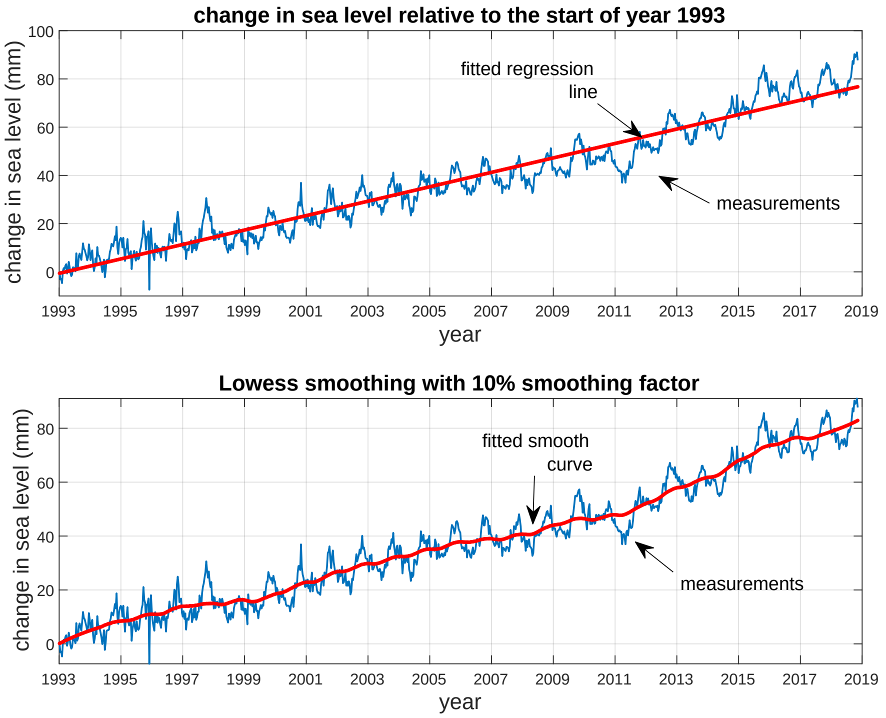

Figure 50.2 (*Top*) Result of fitting a linear regression line onto measurements showing the change in sea level (measured in mm) relative to the start of year 1993. (*Bottom*) Result of fitting a smoother curve to the same data by using the LOWESS procedure described in Example 50.3. The source of the satellite sea level data used in this simulation is from the NASA Goddard Space Flight Center at `https://climate.nasa.gov/vital-signs/sea-level/`.

This construction fits an affine relation (or a line) to the data and allows us to estimate $x(n)$ from an observation $y(n)$ by using (50.60):

$$\widehat{x}(n) = \alpha^\star y(n) - \theta^\star \tag{50.60}$$

The top plot in Fig. 50.2 shows the resulting regression line $\widehat{x}(n)$ along with the measurements $x(n)$ (vertical axis) as a function of the year stamp $y(n)$ (horizontal axis). The bottom plot shows a smoother fitted curve using the LOWESS procedure, which is described next.

Example 50.3 **(LOWESS and LOESS smoothing)** Consider $N-$scalar data pairs denoted by $\{x(n), y(n)\}$, where $n = 0, 1, \ldots, N-1$. In many cases of interest, a regression line is not the most appropriate curve to fit onto the data. We now describe two other popular (but similar) schemes that can be used to fit smoother curves. These schemes are known by the acronyms LOWESS, which stands for "*locally weighted scatter-plot smoothing*" and LOESS, which stands for "*locally estimated scatter-plot smoothing*." Both schemes rely on the use of localized least-squares problems. We describe the LOWESS procedure first.

LOWESS slides a window of width L over the $N-$data points, say, one position at a time. Typical values are $L=N/20, L=N/10$, or $L=N/4$ but other values are possible leading to less (smaller L) or more (larger L) smoothing in the fitted curve. The fraction of samples used within the window is called the smoothing factor, q. Thus, the choice $L=N/10$ corresponds to using $q=10\%$, while the choice $L=N/20$ corresponds to using $q=5\%$. The data in each window are used to estimate one particular point in the window, which is normally (but not always) the middle point. For example, assume we wish to estimate the sample $x(10)$ corresponding to $n=10$, and assume that the window size is $L=5$. In this case, the data samples that belong to the window will be

$$\Big\{(x(8),y(8)),\ (x(9),y(9)),\ \boxed{(x(10),y(10))},\ (x(11),y(11)),\ (x(12),y(12))\Big\} \tag{50.61}$$

with the desired sample $(x(10),y(10))$ appearing at the center of the interval. Clearly, it is not always possible to have the desired sample appear in the middle of the interval. This happens, for example, for the first data point $(x(1),y(1))$. In this case, the other points in the window will lie to its right:

$$\Big\{\boxed{(x(1),y(1))},\ (x(2),y(2)),\ (x(3),y(3)),\ (x(4),y(4)),\ (x(5),y(5))\Big\} \tag{50.62}$$

The same situation happens for the last data point $(x(N-1),y(N-1))$. In this case, the four points in the corresponding window will lie to its left. Regardless, for the data pair $(x(n_o),y(n_o))$ of interest, where we are denoting the index of interest by n_o, we construct a window with L data samples around this point to estimate its $x-$component. For convenience of notation, we collect the indexes of the samples within the window into a set $\mathcal{J}_{n_o}$. For example, for the cases represented in (50.61)–(50.62), we have

$$n_o=10, \qquad \mathcal{J}_{10}=\{8,9,10,11,12\} \tag{50.63}$$

$$n_o=1, \qquad \mathcal{J}_{1}=\{1,2,3,4,5\} \tag{50.64}$$

Let Δ_{n_o} denote the width of the window defined as follows for the above two cases:

$$\Delta_{10}=|y(12)-y(8)|, \qquad \Delta_1=|y(5)-y(1)| \tag{50.65}$$

Next, using the data in each window $\mathcal{J}_{n_o}$, we fit a regression line by solving a *weighted* least-squares problem of the following form:

$$(\alpha^\star_{n_o},\theta^\star_{n_o}) \overset{\Delta}{=} \underset{\alpha,\theta}{\text{argmin}} \left\{\sum_{n\in\mathcal{J}_{n_o}} D(n)\Big(x(n)-(\alpha_{n_o}y(n)-\theta_{n_o})\Big)^2\right\} \tag{50.66}$$

where $D(n)$ is a nonnegative scalar weight constructed as follows:

$$D(n)=\left(1-\left|\frac{y(n)-y(n_o)}{\Delta_{n_o}}\right|^3\right)^3, \quad n\in\mathcal{J}_{n_o} \tag{50.67}$$

Other choices for $D(n)$ are possible, but they need to satisfy certain desirable properties. Observe, for example, that the above choice for the weights varies between 0 and 1, with the weight being equal to 1 at $n=n_o$. Moreover, data samples that are farther away from $y(n_o)$ receive smaller weighting than samples that are closer to it. To solve (50.66), we can again employ the vector notation as follows. We first collect the data from within the window, namely, $\{x(n), y(n)\}_{n\in\mathcal{I}_{n_o}}$, into the vector and matrix quantities:

$$d_{n_o} = \text{col}\Big\{x(n)\Big\}_{n\in\mathcal{I}_{n_o}} \tag{50.68a}$$

$$H_{n_o} = \text{blkcol}\Big\{\,[1\;\, y(n)]\,\Big\}_{n\in\mathcal{I}_{n_o}} \tag{50.68b}$$

$$D_{n_o} = \text{diag}\Big\{D(n)\Big\}_{n\in\mathcal{I}_{n_o}} \tag{50.68c}$$

where D_{n_o} is a diagonal matrix. For example, for the case represented by (50.61) we have

$$d_{10} = \begin{bmatrix} x(8) \\ x(9) \\ x(10) \\ x(11) \\ x(12) \end{bmatrix}, \quad H_{10} = \begin{bmatrix} 1 & y(8) \\ 1 & y(9) \\ 1 & y(10) \\ 1 & y(11) \\ 1 & y(12) \end{bmatrix} \tag{50.69}$$

and

$$D_{10} = \begin{bmatrix} D(8) & & & & \\ & D(9) & & & \\ & & D(10) & & \\ & & & D(11) & \\ & & & & D(12) \end{bmatrix} \tag{50.70}$$

where, for instance,

$$D(11) = \left(1 - \left|\frac{y(11) - y(10)}{y(12) - y(8)}\right|^3\right)^3 \tag{50.71}$$

We also introduce the parameter vector

$$w_{n_o} = \begin{bmatrix} -\theta_{n_o} \\ \alpha_{n_o} \end{bmatrix} \tag{50.72}$$

Then, problem (50.66) is equivalent to

$$w^\star_{n_o} \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^2}{\text{argmin}}\; (d_{n_o} - H_{n_o}w_{n_o})^\mathsf{T} D_{n_o}(d_{n_o} - H_{n_o}w_{n_o}) \tag{50.73}$$

whose solution is given by

$$w^\star_{n_o} = (H^\mathsf{T}_{n_o} D_{n_o} H_{n_o})^{-1} H^\mathsf{T}_{n_o} D_{n_o} d_{n_o} \;\stackrel{\Delta}{=}\; \begin{bmatrix} -\theta^\star_{n_o} \\ \alpha^\star_{n_o} \end{bmatrix} \tag{50.74}$$

This construction now allows us to estimate the sample $x(n_o)$ by using

$$\widehat{x}(n_o) = \alpha^\star_{n_o}\, y(n_o) - \theta^\star_{n_o} \tag{50.75}$$

Next, we slide the window by one position to the right, collect L data points around $(x(n_o+1), y(n_o+1))$ and use them to estimate $x(n_o+1)$ in a similar fashion,

$$\widehat{x}(n_o+1) = \alpha^\star_{n_o+1}\, y(n_o+1) - \theta^\star_{n_o+1} \tag{50.76}$$

and continue in this fashion.

The difference between the LOWESS and LOESS procedures is that the latter fits a second-order curve to the data within each interval $\mathcal{I}_{n_o}$. That is, LOESS replaces (50.66) by

$$(\alpha^\star_{n_o}, \beta^\star_{n_o}, \theta^\star_{n_o}) = \underset{\alpha,\beta,\theta}{\text{argmin}} \left\{ \sum_{n\in\mathcal{I}_{n_o}} D(n)\Big(x(n) - (\alpha_{n_o} y(n) + \beta_{n_o} y^2(n) - \theta_{n_o})\Big)^2 \right\} \tag{50.77}$$

and uses the resulting coefficients $(\alpha^\star_{n_o}, \beta^\star_{n_o}, \theta^\star_{n_o})$ to estimate $x(n_o)$ by using

$$\widehat{x}(n_o) = \alpha^\star_{n_o}\, y(n_o) \;+\; \beta^\star_{n_o} y^2(n_o) - \theta^\star_{n_o} \tag{50.78}$$

We continue to slide the $L-$long window over the data to estimate the subsequent samples $y(n)$.

There is one final step that is normally employed to reduce the effect of outliers that may exist in the data. This step redefines the weights $D(n)$ and repeats the calculation of the first or second-order local curves. Specifically, the following procedure is carried out. Given the target signals $\{x(n)\}$ and the corresponding estimates $\{\widehat{x}(n)\}$ that resulted from the above LOWESS or LOESS construction, we introduce the error sequence

$$e(n) \;\stackrel{\Delta}{=}\; x(n) - \widehat{x}(n), \quad n = 0, 1, 2\ldots, N-1 \tag{50.79}$$

and list the $\{|e(n)|\}$ in increasing order. We then let δ denote the median of this sequence (i.e., the value with as many samples below and above it):

$$\delta \;\stackrel{\Delta}{=}\; \text{median}\{|e(n)|\} \tag{50.80}$$

Using these error quantities, the LOWESS and LOESS implementations introduce the following weighting scalars for $n = 0, 1, \ldots, N-1$:

$$A(n) = \begin{cases} \left(1 - \left|\dfrac{e(n)}{6\delta}\right|^2\right)^2, & \text{if } |e(n)| < 6\delta \\ 0, & \text{otherwise} \end{cases} \tag{50.81}$$

and use them to update $D(n)$ by

$$D(n) \leftarrow D(n)A(n), \quad n \in \mathcal{I}_{n_o} \tag{50.82}$$

We then repeat the design of the local least-squares estimators using these new weights. The construction leads to new estimates $\{\widehat{x}(n)\}$. We can repeat this construction a few times before the process is terminated, leading to the smoothed curve $\{\widehat{x}(n)\}$.

Figure 50.3 shows the LOWESS and LOESS smoothing curves that result from applying the above construction to data measurements representing the change in the global surface temperature (measured in ^{o}C) relative to the average over the period 1951–1980. The data consists of $N = 139$ temperature measurements between the years 1880 and 2018. The top figure shows the curve that results from LOWESS smoothing with a smoothing factor of $q = 5\%$ (corresponding to windows with $L = 6$ samples), while the bottom figure shows the curve that results from LOESS smoothing with a smoothing factor of $q = 10\%$ (corresponding to windows with $L = 13$ samples). Three repeated runs of the form (50.82) are applied.

Example 50.4 (**Confidence levels and interpretability**) One useful feature of least squares solutions is that, under reasonable conditions, we can interpret the results and comment on their confidence level. Consider again the standard least-squares problem (50.19) where we denote the entries of d by $\{x(n)\}$ and the rows of H by $\{h_n^\mathsf{T}\}$, e.g.,

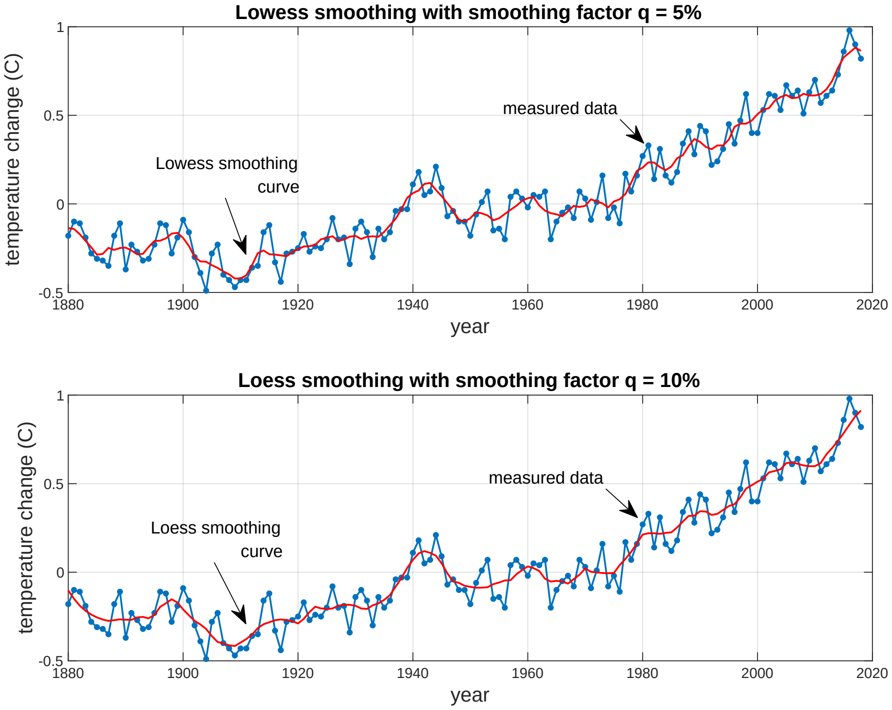

Figure 50.3 LOWESS (*top*) and LOESS (*bottom*) smoothing curves that result from applying the smoothing construction of this example to data measurements representing the change in the global surface temperature (measured in ^{o}C) relative to the average over the period 1951-1980. Three repeated runs of the form (50.82) are applied. The source of the data is the NASA Goddard Institute for Space Studies (GISS) at https://climate.nasa.gov/vital-signs/global-temperature/.

$h_n = \text{col}\{1, y_n\}$ when augmentation is used. When H is full rank, we know that the least-squares solution is given by

$$w^\star = (H^\mathsf{T} H)^{-1} H^\mathsf{T} d \tag{50.83}$$

This vector allows us to predict measurements $x(n)$ using the linear regression model

$$\widehat{x}(n) = h_n^\mathsf{T} w^\star \tag{50.84}$$

There are many ways to assess the quality of the solution in the statistical sciences. We summarize some of the main measures. Using the data $\{x(n)\}$ we define the sample mean and variances:

$$\bar{x} \triangleq \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{50.85a}$$

$$\sigma_x^2 \triangleq \frac{1}{N}\sum_{n=0}^{N-1} (x(n)-\bar{x})^2 \tag{50.85b}$$

$$\sigma_{\widehat{x}}^2 \triangleq \frac{1}{N}\sum_{n=0}^{N-1} (\widehat{x}(n)-\bar{x})^2 \tag{50.85c}$$

$$\sigma_{\widetilde{x}}^2 \triangleq \frac{1}{N}\sum_{n=0}^{N-1} (x(n)-\widehat{x}(n))^2 \tag{50.85d}$$

The variance σ_x^2 measures the squared variation of the samples $x(n)$ around their mean, while the variance $\sigma_{\widehat{x}}^2$ measures the squared variation of the predictions around the same mean. The variance $\sigma_{\widetilde{x}}^2$ measures the squared error between the $x(n)$ and their predictions. It is straightforward to verify that the variance of the target signal decouples into the sum (this is related to the earlier expression (50.40)):

$$\sigma_x^2 = \sigma_{\widehat{x}}^2 + \sigma_{\widetilde{x}}^2 \tag{50.86}$$

The so-called *coefficient of determination* is defined as the ratio:

$$r^2 \triangleq \frac{\sigma_{\widehat{x}}^2}{\sigma_x^2} = 1 - \frac{\sigma_{\widetilde{x}}^2}{\sigma_x^2} \in [0,1] \tag{50.87}$$

This scalar measures the proportion of the variations in $\{x(n)\}$ that is predictable from (or explained by) the observations $\{h_n\}$. For example, if $r = 0.5$, then this means that 25% of the variations in $\{x(n)\}$ can be explained by the variations in $\{h_n\}$. This also means that variations around the regression hyperplane account for 75% of the total variations in the $\{x(n)\}$.

We can assess the quality of the estimated least-squares model $w^\star$ as follows. Assume that the data $\{d, H\}$ satisfy a linear model of the form

$$\boldsymbol{d} = Hw^o + \boldsymbol{v} \tag{50.88}$$

for some unknown $w^o \in \mathbb{R}^M$. The least-squares solution $w^\star$ given by (50.83) is estimating this model. Assume further that $\boldsymbol{v}$ is Gaussian-distributed with $\boldsymbol{v} \sim \mathcal{N}_{\boldsymbol{v}}(0, \sigma_v^2 I_N)$. Then, it is easily seen that $w^\star$ is an unbiased estimator since

$$\begin{aligned} \boldsymbol{w}^\star &= (H^\mathsf{T} H)^{-1} H^\mathsf{T} \boldsymbol{d} \\ &= (H^\mathsf{T} H)^{-1} H^\mathsf{T} (Hw^o + \boldsymbol{v}) \\ &= w^o + (H^\mathsf{T} H)^{-1} H^\mathsf{T}\, \boldsymbol{v} \end{aligned} \tag{50.89}$$

and, consequently, $\mathbb{E}\,\boldsymbol{w}^\star = w^o$. Using the fact that $\boldsymbol{v}$ is Gaussian, we conclude that $\boldsymbol{w}^\star$ is Gaussian-distributed. Its covariance matrix is given by

$$\begin{aligned} \mathbb{E}\,(\boldsymbol{w}^\star - w^o)(\boldsymbol{w}^\star - w^o)^\mathsf{T} &= (H^\mathsf{T} H)^{-1} H^\mathsf{T} \left(\mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\right) H (H^\mathsf{T} H)^{-1} \\ &= \sigma_v^2 (H^\mathsf{T} H)^{-1} \end{aligned} \tag{50.90}$$

In summary, we find that

$$\boldsymbol{w}^\star \sim \mathcal{N}_{\boldsymbol{w}^\star}\Big(w^o, \sigma_v^2 (H^\mathsf{T} H)^{-1}\Big) \tag{50.91}$$

which means that the individual entries of $\boldsymbol{w}^\star$ are Gaussian-distributed with variances given by scaled multiples of the diagonal entries of $(H^{\mathsf{T}}H)^{-1}$. That is, for the $j-$th entry:

$$\boldsymbol{w}^{\star}(j) \;\sim\; \mathbb{N}_{\boldsymbol{w}^{\star}(j)}\Big(w^{o}(j),\; \sigma_v^2\Big[(H^{\mathsf{T}}H)^{-1}\Big]_{jj}\Big) \tag{50.92}$$

in terms of the $j-$th diagonal entry of $(H^{\mathsf{T}}H)^{-1}$. Using this information, we can now determine a 95% confidence interval for each entry $w^{o}(j)$ as follows.

First, we need to introduce the $t-$distribution, also called the Student $t-$distribution. It is symmetric with a similar shape to the Gaussian distribution but has heavier tails. This means that a generic random variable $\boldsymbol{x}$ that is $t-$distributed will have a higher likelihood of assuming extreme values than under a Gaussian distribution. Figure 50.4 compares two Gaussian and $t-$distributions with zero mean and unit variance.

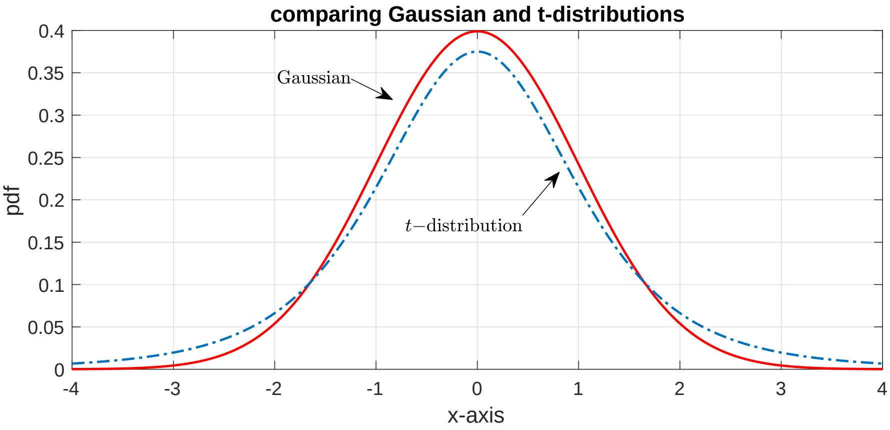

Figure 50.4 Comparing Gaussian and $t-$distributions with zero mean and unit variance. Observe how the $t-$distribution has higher tails.

The $t-$distribution can be motivated as follows. Consider a collection of N scalar independent and identically-distributed realizations arising from a Gaussian distribution with true mean μ and variance σ^2, i.e., $\boldsymbol{x}(n) \sim \mathbb{N}_{\boldsymbol{x}}(\mu, \sigma^2)$. Introduce the sample mean and (unbiased) variance quantities

$$\bar{x} \;\stackrel{\Delta}{=}\; \frac{1}{N}\sum_{n=1}^{N} x(n), \qquad s_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x(n)-\bar{x})^2 \tag{50.93}$$

The quantities $\{\bar{x}, s_x^2\}$ should be viewed as random variables, written in boldface notation $\{\bar{\boldsymbol{x}}, \boldsymbol{s}_x^2\}$, because their values vary with the randomness in selecting the $\{\boldsymbol{x}(n)\}$. Next, we define the $t-$score variable, which measures how far the sample mean is from the true mean (scaled by the sample standard deviation and $\sqrt{N}$):

$$\boldsymbol{t} \;\stackrel{\Delta}{=}\; \frac{\bar{\boldsymbol{x}}-\mu}{\boldsymbol{s}_x/\sqrt{N}} \tag{50.94}$$

The pdf of the $\boldsymbol{t}$ variable is called the $t-$distribution with $d = N-1$ degrees of freedom. It has zero mean and unit variance and is formally defined by the expression:

$$f_{\boldsymbol{t}}(t;d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\,\frac{1}{\sqrt{d\pi}}\,\frac{1}{(1+t^2/d)^{(d+1)/2}}, \qquad (\boldsymbol{t}-\textbf{distribution}) \tag{50.95}$$

where $\Gamma(x)$ refers to the Gamma function encountered earlier in Prob. 4.3. The definition (50.94) explains why the $t-$distribution is useful in constructing confidence intervals. That is because it assesses how the sample mean is distributed around the true mean. Due to its relevance, the $t-$distribution appears tabulated in many texts on statistics and these tables are used in the following manner.

Let $\alpha = 5\%$ (this value is known as the desired *significance level* in statistics). We use a table of $t-$distributions to determine the *critical value* denoted by $t^{N-M}_{\alpha/2}$; this is the value in a $t-$distribution with $N - M$ degrees of freedom beyond which the area under the pdf curve will be 2.5% (this calculation amounts to performing what is known as a *one-tailed test*) — see Fig. 50.5. An example of this tabular form is shown in Table 50.1. One enters the value of $\alpha/2$ along the vertical direction and the degree $N - M$ along the horizontal direction and reads out the entry corresponding to $t^{N-M}_{\alpha/2}$. For example, using $N - M = 15$ degrees of freedom and $\alpha/2 = 2.5\%$, one reads the value marked in bold face $t^{15}_{2.5\%} = 2.131$.

Table 50.1 Critical values of $t^{d}_{\alpha/2}$ in one-tailed t-tests with d degrees of freedom. The values in the last row can be used for large degrees of freedom.

degree d	5%	**2.5%**	1%	0.5%	0.1%
1	6.314	12.706	31.821	63.657	318.309
2	2.920	4.303	6.965	9.925	22.327
3	2.353	3.182	4.541	5.841	10.215
4	2.132	2.776	3.747	4.604	7.173
5	2.015	2.571	3.365	4.032	5.893
6	1.943	2.447	3.143	3.707	5.208
7	1.894	2.365	2.998	3.499	4.785
8	1.860	2.306	2.896	3.355	4.501
9	1.833	2.262	2.821	3.250	4.297
10	1.812	2.228	2.764	3.169	4.144
11	1.796	2.201	2.718	3.106	4.025
12	1.782	2.179	2.681	3.055	3.930
13	1.771	2.160	2.650	3.012	3.852
14	1.761	2.145	2.624	2.977	3.787
15	1.753	**2.131**	2.602	2.947	3.733
16	1.746	2.120	2.583	2.921	3.686
17	1.740	2.110	2.567	2.898	3.646
18	1.734	2.101	2.552	2.878	3.610
19	1.729	2.093	2.539	2.861	3.579
20	1.725	2.086	2.528	2.845	3.552
∞	1.645	1.960	2.326	2.576	3.090

Once $t^{N-M}_{\alpha/2}$ is determined, the confidence interval for each entry of w^o would be given by

$$w^\star(j) \;\pm\; t^{N-M}_{\alpha/2}\,\sigma_v\,\sqrt{[(H^\mathsf{T} H)^{-1}]_{jj}} \tag{50.96}$$

This means that there is 95% chance that the true value $w^o(j)$ lies within the interval.

Likewise, given an observation h_n, we can derive a confidence interval for the unperturbed component $h_n^\mathsf{T} w^o$, which happens to be the mean of $\boldsymbol{x}(n)$ in model (50.88). That

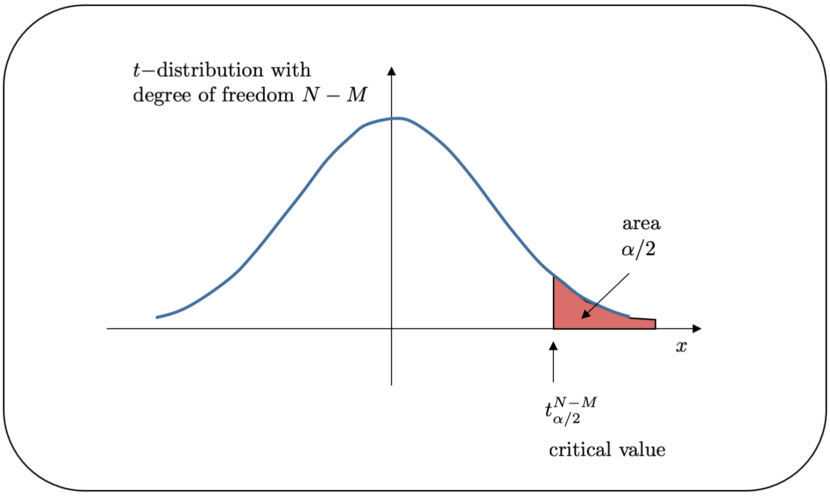

Figure 50.5 The critical value $t_{\alpha/2}^{N-M}$ is the point to the right of which the area under a $t-$distribution with degree $N-M$ is equal to $\alpha/2$.

is, we can derive a confidence interval for the *expected* value of the target signal $\boldsymbol{x}(n)$ that would result from h_n. To see this, consider the prediction $\widehat{\boldsymbol{x}}(n) = h_n^{\mathsf{T}}\boldsymbol{w}^\star$. This prediction is again Gaussian-distributed since $\boldsymbol{w}^\star$ is Gaussian. Its mean and variance are found as follows. First note that

$$\begin{aligned}\widehat{\boldsymbol{x}}(n) &= h_n^{\mathsf{T}}\,\boldsymbol{w}^\star\\ &= h_n^{\mathsf{T}}\Big\{w^o + (H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\,\boldsymbol{v}\Big\}\\ &= h_n^{\mathsf{T}}w^o + h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\boldsymbol{v}\end{aligned} \tag{50.97}$$

We conclude that $\mathbb{E}\,\widehat{\boldsymbol{x}}(n) = h_n^{\mathsf{T}}w^o$, so that the mean of the prediction agrees with the actual mean, $\mathbb{E}\,\boldsymbol{x}(n) = h_n^{\mathsf{T}}w^o$. Moreover, the prediction variance is given by

$$\begin{aligned}\mathbb{E}\,(\widehat{\boldsymbol{x}}(n) - h_n^{\mathsf{T}}w^o)^2 &= h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\,(\mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}})\,H(H^{\mathsf{T}}H)^{-1}h_n\\ &= \sigma_v^2 h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n\end{aligned} \tag{50.98}$$

so that

$$\widehat{\boldsymbol{x}}(n) \;\sim\; \mathcal{N}_{\widehat{\boldsymbol{x}}(n)}\Big(h_n^{\mathsf{T}}w^o,\; \sigma_v^2 h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n\Big) \tag{50.99}$$

which shows that the predictions will be Gaussian-distributed around the actual mean, $h_n^{\mathsf{T}}w^o$. We can then determine a 95% confidence interval for the mean value $h_n^{\mathsf{T}}w^o$ by using

$$\widehat{x}(n) \;\pm\; t_{\alpha/2}^{N-M}\,\sigma_v\,\sqrt{h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n} \tag{50.100}$$

Given an observation h_n, this means that there is 95% chance that the mean value $h_n^{\mathsf{T}}w^o$ will lie within the above interval around $\widehat{x}(n)$.

In a similar vein, given a feature h_n, we can derive a confidence interval for the target $\boldsymbol{x}(n)$ itself (rather than its mean, as was done above). To see this, we note that the difference $\widehat{\boldsymbol{x}}(n) - \boldsymbol{x}(n)$ is again Gaussian distributed, albeit with mean zero since

$$\begin{aligned}\widehat{\boldsymbol{x}}(n) - \boldsymbol{x}(n) &= \Big(h_n^{\mathsf{T}} w^o + h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\boldsymbol{v}\Big) - (h_n^{\mathsf{T}} w^o + \boldsymbol{v}(n)) \\ &= h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\boldsymbol{v} - \boldsymbol{v}(n) \end{aligned} \tag{50.101}$$

Moreover, the variance is given by

$$\mathbb{E}\,(\,\widehat{\boldsymbol{x}}(n) - \boldsymbol{x}(n))^2 = \sigma_v^2(1 - h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n) \tag{50.102}$$

so that

$$\widehat{\boldsymbol{x}}(n) \;\sim\; \mathcal{N}_{\widehat{\boldsymbol{x}}(n)}\Big(x(n),\; \sigma_v^2(1 - h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n)\Big) \tag{50.103}$$

This result shows that the predictions will be Gaussian-distributed around the actual value $x(n)$. We can then determine a 95% confidence interval for $x(n)$ by using

$$\widehat{x}(n) \;\pm\; t_{\alpha/2}^{N-M}\,\sigma_v\,\sqrt{1 - h_n^{\mathsf{T}}(H^{\mathsf{T}}H)^{-1}h_n} \tag{50.104}$$

The expressions so far assume knowledge of σ_v^2. If this information is not available, it can be estimated by noting that $v(n) = x(n) - h_n^{\mathsf{T}} w^o$ and using the sample approximation:

$$\widehat{\sigma}_v^2 \approx \frac{1}{N-1}\sum_{n=0}^{N-1}\Big(x(n) - h_n^{\mathsf{T}} w^\star\Big)^2 \tag{50.105}$$

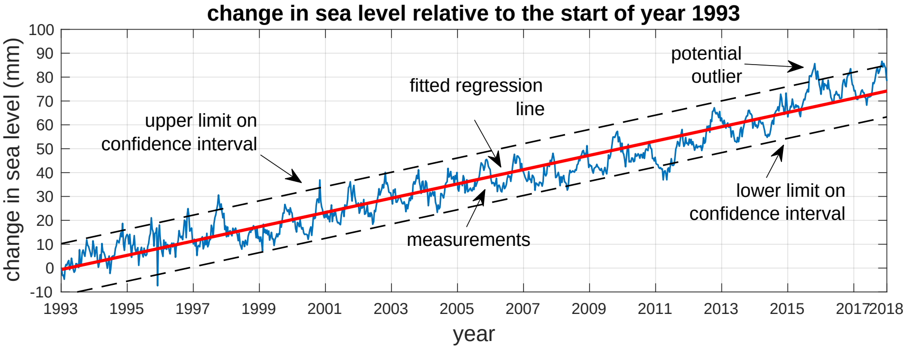

Figure 50.6 The fitted regression line is shown in solid red color, while the lines that correspond to the upper and lower limits of the confidence interval (50.104) appear in dotted format.

The analysis in this example is meant to illustrate that, for least-squares problems and under some reasonable conditions, we are able to assess the confidence levels we have in the results. This is a useful property for learning algorithms to have so that their results become amenable to a more judicious interpretation. It also enables the algorithms to detect outliers and malicious data. For example, if some data pair $(x(m), h_m)$ is received, one may compute $\widehat{x}(m) = h_m^{\mathsf{T}} w^\star$ and verify whether $x(m)$ lies within the corresponding confidence interval (constructed according to (50.104) with n replaced by m). If not, then one can flag this data point as being an outlier.

We apply construction (50.104)–(50.105) to Example 50.2, which involved fitting a regression line to sea levels over multiple years. We use $N = 952$ and $M = 2$ (due to the augmentation of the feature data by the unit entry) so that the number of degrees of freedom is 950. Using the data from the last row of Table 50.1 we have $t_{2.5\%}^{950} \approx 1.960$.

The regression lines that result from using the lower and upper limits in (50.104) appear in dotted format in Fig. 50.6.

Example 50.5 (**Sketching**) In big data applications, the amount of available data can be massive, giving rise to situations where $N \gg M$, i.e., the number of observations far exceeds the number of unknowns in the least-squares problem (50.19). In these cases, the solution of the normal equations (50.25) becomes prohibitively expensive since computing the products $H^{\mathsf{T}}H$ and $H^{\mathsf{T}}d$ require $O(NM^2)$ and $O(NM)$ additions and multiplications, respectively. One technique to reduce the computational complexity is to employ *randomized algorithms* that rely on the concept of *sketching*. The purpose of these algorithms is to seek approximate solutions, denoted by w^s, with the following useful property: with high probability $1-\delta$, the solution w^s should lead to a risk value that is $\epsilon-$close to the optimal risk value, namely, it should hold that:

$$\mathbb{P}\Big(\|d - Hw^s\|^2 \;\leq\; (1+\epsilon)\|d - Hw^\star\|^2\Big) = 1 - \delta \tag{50.106}$$

where $\delta > 0$ is a small positive number. Sketching procedures operate as follows. They first select some *random* matrix S of size $R \times N$, with $R \ll N$. Subsequently, they compute the products Sd and SH and determine w^s by solving the altered least-squares problem:

$$w^s \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}}\;\; \|Sd - SHw\|^2 \tag{50.107}$$

Observe that this is a smaller-size problem because SH is now $R \times M$. Since there is some nonzero probability of failure in (50.106), it is customary to repeat the sketching construction several times (by choosing different sketching matrices S each time), and then keep the best solution w^s from among the repeated experiments (i.e., the one with the smallest risk value).

The three main challenges that arise in sketching solutions relate to: **(a)** selecting sketching matrices S that guarantee (50.106), **(b)** selecting suitable values for the dimension R, and, more importantly, **(c)** choosing sketching matrices S for which the products Sd and SH can be computed efficiently. For this last condition, it is desirable to seek *sparse* choices for S.

One option is to employ *Gaussian sketching*. We select a dimension $R = O((M \log M)/\epsilon)$ and let the entries of S be independent and identically distributed Gaussian random variables with zero mean and variance equal to $1/R$. This construction can be shown to answer points (a) and (b) above, but is costly to implement since it generally leads to dense matrices S for which point (c) is expensive. Computing SH in this case requires $O(NM^2 \log M)$ computations.

A second option that also answers points (a) and (b) above is to employ a random subsampling strategy as follows. We introduce the singular value decomposition of H (this is of course a costly step and that is the reason why this option will not be viable in general):

$$H = U_H \Sigma_H V_H^{\mathsf{T}} \tag{50.108}$$

where U_H is $N \times N$ orthonormal; its rows have N entries each. We let u_n^{T} denote the *restriction* of the $n-$th row to its M leading entries. That is, each u_n^{T} consists of the first M entries in the $n-$th row of U_H. The so-called *leverage scores* of H are defined as the squared norms of these restricted vectors:

$$\ell_n = \|u_n\|^2, \quad n = 1, 2, \ldots, N \tag{50.109}$$

It is straightforward to verify that the leverage scores correspond to the diagonal entries of the projection matrix onto $\mathcal{R}(H)$, namely,

$$\ell_n = \big[\mathcal{P}_H\big]_{nn}, \quad n = 1, 2, \ldots, N \tag{50.110}$$

We normalize the leverage scores by dividing by their sum to define a probability distribution over the integer indexes $1 \leq r \leq N$:

$$p_n \overset{\Delta}{=} \mathbb{P}(\boldsymbol{r} = n) = \ell_n \Big/ \sum_{m=1}^{N} \ell_m \tag{50.111}$$

The scalar p_n defines the probability of selecting at random the index value n. Next, for each row $r = 1, 2, \ldots, R$ of the sketching matrix S:

(a) We select an index n at random from the interval $\{1, 2, \ldots, N\}$ with probability equal to p_n.

(b) We set the $r-$th row of S to the basis vector e_n^{T} scaled by $1/\sqrt{Rp_n}$, where $e_n \in \mathbb{R}^N$ has a unit entry at the $n-$th location and zeros elsewhere.

Observe that, under this construction, each row of S will contain a single unit entry. In this way, the multiplication of this row by H ends up selecting a row from H. For this reason, we refer to S as performing *random subsampling*. The main inconvenience of this construction is that it requires computation of the leverage scores, which in turn require knowledge of the SVD factor U_H. It would be useful to seek sketching matrices that are data-independent.

The third construction achieves this goal and is based on selecting a random subsampling Hadamard matrix. Assume $N = 2^n$ (i.e., N is a power of 2) and select $R = O((M\log^3 N)/\epsilon)$. Introduce the $N \times N$ orthonormal Hadamard matrix computed as the Kronecker product of 2×2 orthonormal Hadamard matrices:

$$\mathcal{H} = \underbrace{\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right) \otimes \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right) \otimes \ldots \otimes \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right)}_{n \text{ times}} \tag{50.112}$$

Apart from scaling by $1/(\sqrt{2})^n = 1/\sqrt{N}$, the entries of $\mathcal{H}$ will be ±1. Next, we

(a) Select uniformly at random R rows from $\mathcal{H}$, and denote the resulting $R \times N$ matrix by $\mathcal{H}_R$;

(b) Construct an $N \times N$ random sign matrix in the form of a diagonal matrix D with random ± 1 entries on its diagonal, with each entry selected with probability 1/2;

(c) And then set $S = \sqrt{N/R}\,\mathcal{H}_R D$.

It can be verified that under this third construction, the complexity of determining w^s is $O(NM\log(N/\epsilon) + (M^3\log^3 N)/\epsilon)$. The purpose of this example is to introduce the reader to the concept of sketching in the context of least-squares problems. Additional comments are provided at the end of the chapter.

50.3 RECURSIVE LEAST-SQUARES

One key advantage of the least-squares empirical risk (50.16) is that it enables an *exact* recursive computation of the minimizer. The recursive solution is particularly useful for situations involving streaming data arriving successively over time.

In this section we derive the recursive least-squares (RLS) algorithm but first introduce two modifications into the empirical risk function for two main reasons: **(a)** to enable an exact derivation of the recursive solution, and **(b)** to incorporate a useful tracking mechanism into the algorithm.

50.3.1 Exponential Weighting

We modify the least-squares empirical risk (50.16) to the following exponentially weighted form with ℓ_2-regularization:

$$\min_{w\in\mathbb{R}^M} \left\{\frac{1}{N}\rho'\lambda^N\|w\|^2 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\lambda^{N-1-n}\left(x(n)-y_n^{\mathsf{T}}w\right)^2\right\} \tag{50.113}$$

There are three modifications in this formulation, which we motivate as follows:

(a) (**Exponential weighting**). The scalar $0 \ll \lambda < 1$ is called the *forgetting factor* and is a number close to one. Its purpose is to scale down data from the past more heavily than recent data. For example, in the above risk, data from time $n = 0$ is scaled by λ^{N-1} while data at $n = N-1$ is scaled by one. In this way, the algorithm is endowed with a memory mechanism that "forgets" older data and emphasizes recent data. This is a useful property to enable the algorithm to track drifts in the statistical properties of the data, especially when the subscript n has a time connotation and is used to index streaming data. The special case $\lambda = 1$ is known as *growing memory.* Exponential weighting is one form of data windowing where the effective length of the window is approximately $1/(1-\lambda)$ samples.

(b) (**Decaying ℓ_2-regularization**). The scalar $\rho' > 0$ is an ℓ_2-regularization parameter. Observe though that the penalty term $\rho'\|w\|^2$ in (50.113) is scaled by λ^N as well; this factor dies out with time at an exponential rate and helps eliminate regularization after sufficient data have been processed. In other words, regularization will be more pronounced during the initial stages of the recursive algorithm and less pronounced later. One advantage of the regularization factor is that it helps ensure that the coefficient matrix that is inverted in future expression (50.121b) is nonsingular.

(c) (**Sample averaging**). In addition, both terms in (50.113) are scaled by $1/N$, which is independent of w. For this reason, we can ignore the $1/N$ factor and solve instead:

$$w_{N-1} \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{\rho'\lambda^N\|w\|^2 \;+\; \sum_{n=0}^{N-1}\lambda^{N-1-n}\left(x(n)-y_n^{\mathsf{T}}w\right)^2\right\} \tag{50.114}$$

where we are now denoting the unique solution by w_{N-1} rather than $w^\star$. The subscript $N-1$ is meant to indicate that the solution w_{N-1} is based on data up to time $N-1$. We attach the time subscript to the solution because we will be deriving a recursive construction that allows us to compute w_N

from w_{N-1} where w_N is minimizes the enlarged risk:

$$w_N \stackrel{\Delta}{=} \underset{w\in{\rm I\!R}^M}{\text{argmin}} \left\{\rho'\lambda^{N+1}\|w\|^2 \;+\; \sum_{n=0}^{N}\lambda^{N-n}\left(x(n)-y_n^{\mathsf T}w\right)^2\right\} \tag{50.115}$$

where a new pair of data, $\{x(N), y_N\}$, has been added to the risk. The adjustments introduced through steps (b) and (c) enable the derivation of an exact recursive algorithm, as the argument will show.

In a manner similar to (50.17), we introduce the data quantities:

$$H_N \stackrel{\Delta}{=} \begin{bmatrix} y_0^{\mathsf T}\\ y_1^{\mathsf T}\\ y_2^{\mathsf T}\\ \vdots\\ y_N^{\mathsf T}\end{bmatrix}, \quad d_N \stackrel{\Delta}{=} \begin{bmatrix} x(0)\\ x(1)\\ x(2)\\ \vdots\\ x(N)\end{bmatrix} \tag{50.116}$$

where we are now attaching a time subscript to $\{H_n, d_N\}$ to indicate that they involve data up to time N. Thus, note that we can partition them in the form:

$$H_N = \left[\begin{array}{c} H_{N-1}\\ \hline y_N^{\mathsf T}\end{array}\right], \quad d_N = \left[\begin{array}{c} d_{N-1}\\ \hline x(N)\end{array}\right] \tag{50.117}$$

so that $\{H_{N-1}, H_N\}$ differ by one row and $\{d_{N-1}, d_N\}$ differ by one entry. We also introduce the diagonal weighting matrix:

$$\Lambda_N \stackrel{\Delta}{=} \text{diag}\left\{\lambda^N,\, \lambda^{N-1}, \ldots,\, 1\right\} \tag{50.118}$$

and note that

$$\Lambda_N = \begin{bmatrix} \lambda\Lambda_{N-1} & \\ & 1\end{bmatrix} \tag{50.119}$$

Using $\{H_n, d_N, \Lambda_N\}$, problems (50.114) and (50.115) can be rewritten in matrix form as follows:

$$w_{N-1} \stackrel{\Delta}{=} \underset{w\in{\rm I\!R}^M}{\text{argmin}} \left\{\rho'\lambda^N\|w\|^2 \;+\; (d_{N-1}-H_{N-1}w)^{\mathsf T}\Lambda_{N-1}(d_{N-1}-H_{N-1}w)\right\} \tag{50.120a}$$

$$w_N \stackrel{\Delta}{=} \underset{w\in{\rm I\!R}^M}{\text{argmin}} \left\{\rho'\lambda^{N+1}\|w\|^2 \;+\; (d_N-H_Nw)^{\mathsf T}\Lambda_N(d_N-H_Nw)\right\} \tag{50.120b}$$

Differentiating the above risks relative to w, we find that the unique solutions w_N and w_{N-1} are given by the expressions:

$$w_{N-1} = \left(\rho'\lambda^N I_M + H_{N-1}^{\mathsf T}\Lambda_{N-1}H_{N-1}\right)^{-1} H_{N-1}^{\mathsf T}\Lambda_{N-1}d_{N-1} \tag{50.121a}$$

$$w_N = \left(\rho'\lambda^{N+1} I_M + H_N^{\mathsf T}\Lambda_N H_N\right)^{-1} H_N^{\mathsf T}\Lambda_N d_N \tag{50.121b}$$

These equations allow us to evaluate the solutions $\{w_{N-1}, w_N\}$ directly from the data matrices. However, a more efficient construction is possible by going from w_{N-1} to w_N more directly as we explain next. This step will be referred to as the *time-update* step.

50.3.2 Exponentially-Weighted RLS

To derive the recursive algorithm, we introduce the following three quantities:

$$P_N \stackrel{\Delta}{=} \left(\rho'\lambda^{N+1}I_M + H_N^{\mathsf{T}}\Lambda_N H_N\right)^{-1} \tag{50.122a}$$

$$t(N) \stackrel{\Delta}{=} 1/(1+\lambda^{-1}y_N^{\mathsf{T}}P_{N-1}y_N) \tag{50.122b}$$

$$g_N \stackrel{\Delta}{=} \lambda^{-1}P_{N-1}y_N t(N) \tag{50.122c}$$

where P_N is $M\times M$, g_N is an $M\times 1$ gain vector, and $t(N)$ is a scalar factor. The derivation below establishes the following result. Given $\rho' > 0$ and a forgetting factor $0 \ll \lambda \leq 1$, the solution w_N of the exponentially-weighted regularized least-squares problem (50.115), and the corresponding minimum risk denoted by $\xi(N)$, can be computed recursively as shown in listing (50.123) — see Prob. 50.7 for a derivation of the recursion for the minimum cost.

Recursive least-squares (RLS) for solving (50.115)

```
given $N$ data pairs $\{x(n) \in \mathbb{R}, y_n \in \mathbb{R}^M\}$, $n = 0, 1, \ldots, N-1$;
start with $P_{-1} = \frac{1}{\rho'}I_M$, $\xi(-1) = 0, w_{-1} = 0_M$;
repeat for $n = 0, 1, 2, \ldots, N-1$:
    $t(n) = 1/(1+\lambda^{-1}y_n^{\mathsf{T}}P_{n-1}y_n)$
    $g_n = \lambda^{-1}P_{n-1}y_n t(n)$
    $\widehat{x}(n) = y_n^{\mathsf{T}}w_{n-1}$
    $e(n) = x(n) - \widehat{x}(n)$
    $w_n = w_{n-1} + g_n e(n)$
    $P_n = \lambda^{-1}P_{n-1} - g_n g_n^{\mathsf{T}}/t(n)$
    $\xi(n) = \lambda\xi(n-1) + t(n)e^2(n)$
end
```
(50.123)

Derivation of (50.123) We first rewrite (50.121a)–(50.121b) more compactly using the matrices $\{P_{N-1}, P_N\}$ as:

$$w_{N-1} = P_{N-1}H_{N-1}^{\mathsf{T}}\Lambda_{N-1}d_{N-1} \tag{50.124a}$$

$$w_N = P_N H_N^{\mathsf{T}}\Lambda_N d_N \tag{50.124b}$$

Next, we exploit the relations between $\{H_N, d_N, \Lambda_N\}$ and $\{H_{N-1}, d_{N-1}, \Lambda_{N-1}\}$ from

(50.117) and (50.119) in order to relate w_{N-1} to w_N directly. To begin with, note that

$$\begin{aligned}
P_N^{-1} &= \rho'\lambda^{N+1}I_M \;+\; H_N^{\mathsf T}\Lambda_N H_N \\
&\overset{(50.119)}{=} \rho'\lambda\lambda^N I_M \;+\; \lambda H_{N-1}^{\mathsf T}\Lambda_{N-1}H_{N-1} + y_N y_N^{\mathsf T} \\
&= \lambda P_{N-1}^{-1} + y_N y_N^{\mathsf T} \qquad (50.125)
\end{aligned}$$

Then, by using the matrix inversion identity (29.89) with the identifications

$$A \leftarrow \lambda P_{N-1}^{-1}, \qquad B \leftarrow y_N, \qquad C \leftarrow 1, \qquad D \leftarrow y_N^{\mathsf T} \qquad (50.126)$$

we obtain a recursive formula for updating P_N directly rather than its inverse,

$$P_N = \lambda^{-1}P_{N-1} - \frac{\lambda^{-1}P_{N-1}y_N y_N^{\mathsf T}P_{N-1}\lambda^{-1}}{1+\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}, \qquad P_{-1} = \frac{1}{\rho'}I_M \qquad (50.127)$$

This recursion for P_N also gives one for updating the regularized solution w_N itself. Using expression (50.124b) for w_N, and substituting the above recursion for P_N, we find

$$\begin{aligned}
w_N &= P_N\Big(\lambda H_{N-1}^{\mathsf T}\Lambda_{N-1}d_{N-1} + y_N x(n)\Big) \\
&\overset{(50.127)}{=} \left(\lambda^{-1}P_{N-1} - \frac{\lambda^{-1}P_{N-1}y_N y_N^{\mathsf T}P_{N-1}\lambda^{-1}}{1+\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}\right)\Big(\lambda H_{N-1}^{\mathsf T}\Lambda_{N-1}d_{N-1} + y_N x(n)\Big) \\
&= \underbrace{P_{N-1}H_{N-1}^{\mathsf T}\Lambda_{N-1}d_{N-1}}_{=w_{N-1}} - \frac{\lambda^{-1}P_{N-1}y_N}{1+\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}y_N^{\mathsf T}\underbrace{P_{N-1}H_{N-1}^{\mathsf T}\Lambda_{N-1}d_{N-1}}_{=w_{N-1}} \\
&\quad + \lambda^{-1}P_{N-1}y_N\left(1 - \frac{\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}{1+\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}\right)x(n) \qquad (50.128)
\end{aligned}$$

That is,

$$w_N = w_{N-1} + \frac{\lambda^{-1}P_{N-1}y_N}{1+\lambda^{-1}y_N^{\mathsf T}P_{N-1}y_N}\big(x(n) - y_N^{\mathsf T}w_{N-1}\big), \qquad w_{-1} = 0 \qquad (50.129)$$

■

The RLS implementation (50.123) updates the weight iterate from w_{n-1} to w_n for each data pair $\{x(n), y_n\}$. Such implementations are useful for situations involving *streaming* data where one data pair arrives at each time instant n and the algorithm responds to it by updating w_{n-1} to w_n in real-time. If desired, we can extend the algorithm to deal with blocks of data as explained in Prob. 50.30.

50.3.3 Useful Relations

The scalar $t(n)$ in algorithm is called the "*conversion factor.*" This is because it transforms *a-priori* errors into *a-posteriori* errors, as established in Prob. 50.17. Some straightforward algebra, using recursion (50.127) for P_n, shows that $\{g_n, t(n)\}$ can also be expressed in terms of P_n, namely,

$$g_n = P_n y_n \qquad (50.130a)$$

$$t(n) = 1 - y_n^{\mathsf T}g_n \qquad (50.130b)$$

To justify (50.130a)–(50.130b), we simply note the following. Multiplying recursion (50.127) for P_n by y_n from the right we get

$$\begin{aligned} P_n y_n &= \lambda^{-1} P_{n-1} y_n - \frac{\lambda^{-1} P_{n-1} y_n y_n^{\mathsf{T}} P_{n-1} y_n \lambda^{-1}}{1 + \lambda^{-1} y_n^{\mathsf{T}} P_{n-1} y_n} \\ &= \frac{\lambda^{-1} P_{n-1} y_n}{1 + \lambda^{-1} y_n^{\mathsf{T}} P_{n-1} y_n} \\ &= g_n \end{aligned} \tag{50.131}$$

By further multiplying the above identity by y_n^{T} from the left we get

$$y_n^{\mathsf{T}} P_n y_n = \frac{\lambda^{-1} y_n^{\mathsf{T}} P_{n-1} y_n}{1 + \lambda^{-1} y_n^{\mathsf{T}} P_{n-1} y_n} \tag{50.132}$$

so that, by subtracting 1 from both sides, we obtain (50.130b).

Furthermore, we note that at each iteration n, the variable P_n in the algorithm is equal to the following quantity:

$$P_n = \left(\rho' \lambda^{n+1} I_M + H_n^{\mathsf{T}} \Lambda_n H_n\right)^{-1} \tag{50.133}$$

and the iterate w_n is the solution to the regularized least-squares problem that uses only the data data up to time n:

$$w_n \overset{\Delta}{=} \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \left\{ \rho' \lambda^{n+1} \|w\|^2 + \sum_{m=0}^{n} \lambda^{n-m} \Big(x(m) - y_m^{\mathsf{T}} w \Big)^2 \right\} \tag{50.134}$$

The minimum cost for this problem, with w replaced by w_n, is equal to $\xi(n)$.

Example 50.6 **(Recommender systems)** We revisit the recommender system studied earlier in Example 16.7. There we introduced a collaborative filtering approach based on matrix factorization to predict ratings by users. We denoted the weight vector by user u by $w_u \in \mathbb{R}^M$ and the latent feature vector for item i by $h_i \in \mathbb{R}^M$. Subsequently, we formulated the regularized least-squares optimization problem:

$$\begin{aligned} \left\{ \widehat{w}_u, \widehat{h}_i, \widehat{\theta}_u, \widehat{\alpha}_i \right\} = \underset{\{w_u, h_i, \theta_u, \alpha_i\}}{\operatorname{argmin}} & \left\{ \sum_{u=1}^{U} \rho \|w_u\|^2 + \sum_{i=1}^{I} \rho \|h_i\|^2 + \right. \\ & \left. \sum_{(u,i) \in \mathcal{R}} \Big(r_{ui} - h_i^{\mathsf{T}} w_u + \theta_u + \alpha_i \Big)^2 \right\} \end{aligned} \tag{50.135}$$

where the last sum is over the valid indexes $(i, u) \in \mathcal{R}$, i.e., over the indexes for which valid ratings exist. All entries with missing ratings are therefore excluded. We approximated the minimizer of the above (non-convex) problem by applying the stochastic gradient solution (16.58). In this example, we pursue instead an *alternating least-squares solution.*

Note that if we fix any three of the parameters, then the risk function is quadratic over the remaining parameter. For example, if we fix $(h_i, \theta_u, \alpha_i)$, then the risk is quadratic over w_u. For any index u, let the notation $\mathcal{R}_u$ represent the set of valid indexes i for which (u, i) has a rating. Note that u is fixed within $\mathcal{R}_u$. Likewise, for any index i, let the notation $\mathcal{R}_i$ represent the set of valid indexes u for which (u, i) has a rating. Note that i is fixed within $\mathcal{R}_i$.

For any specific u, setting the gradient relative to w_u to zero leads to the expression:

$$\widehat{w}_u = \left(\sum_{i\in\mathcal{R}_u}(\rho I_M + h_i h_i^{\mathsf{T}})\right)^{-1}\left(\sum_{i\in\mathcal{R}_u} h_i(r_{ui}+\theta_u+\alpha_i)\right) \tag{50.136}$$

We can obtain similar expressions for $\widehat{h}_i$, $\widehat{\theta}_u$ and $\widehat{\alpha}_i$, leading to listing (50.137). In the listing, the term $w_{u,m}$ represents the estimate for w_u at iteration m; likewise for $h_{i,m}$, $\theta_u(m)$, and $\alpha_i(m)$.

Alternating least-squares algorithm for problem (50.135)

```
given ratings $r_{u,i}$ for $(u,i)\in\mathcal{R}$;
start with arbitrary $\{\boldsymbol{w}_{u,-1}, \boldsymbol{h}_{i,-1}, \boldsymbol{\theta}_u(-1), \boldsymbol{\alpha}_i(-1)\}$;
repeat until convergence over $m=0,1,\ldots$:
    repeat for $u=1,\ldots,U$ :
        $A_u = \sum_{i\in\mathcal{R}_u}(\rho I_M + h_{i,m-1}h_{i,m-1}^{\mathsf{T}})$
        $w_{u,m} = A_u^{-1}\left(\sum_{i\in\mathcal{R}_u} h_{i,m-1}(r_{ui}+\theta_u(m-1)+\alpha_i(m-1))\right)$
        $\theta_u(m) = -\frac{1}{|\mathcal{R}_u|}\sum_{i\in\mathcal{R}_u}\Big(r_{ui} - h_{i,m-1}^{\mathsf{T}}w_{u,m-1} + \alpha_i(m-1)\Big)$
    end
    repeat for $i=1,\ldots,I$ :
        $B_i = \sum_{u\in\mathcal{R}_i}(\rho I_M + w_{u,m}w_{u,m}^{\mathsf{T}})$
        $h_{i,m} = B_i^{-1}\left(\sum_{u\in\mathcal{R}_i} w_{u,m}(r_{ui}+\theta_u(m)+\alpha_i(m-1))\right)$
        $\alpha_i(m) = -\frac{1}{|\mathcal{R}_i|}\sum_{u\in\mathcal{R}_i}\Big(r_{ui} - h_{i,m-1}^{\mathsf{T}}w_{u,m} + \theta_u(m)\Big)$
    end
end
return $\{w_u^\star, h_i^\star, \theta_u^\star, \alpha_i^\star\}$
```
(50.137)

We simulate recursions (50.137) for the same situation discussed earlier in Example 16.7. We consider the same ranking matrix for $U = 10$ users and $I = 10$ items with integer scores in the range $1 \leq r \leq 5$; unavailable scores are marked by the symbol ?:

$$R = \begin{bmatrix} 5 & 3 & 2 & 2 & ? & 3 & 4 & ? & 3 & 3 \\ 5 & 4 & 1 & 3 & 1 & 4 & 4 & ? & 3 & ? \\ 3 & 5 & ? & 2 & 1 & 5 & 4 & 1 & 4 & 1 \\ ? & 2 & 3 & 4 & 4 & 5 & 2 & 5 & 1 & 1 \\ 2 & 1 & 2 & 2 & 1 & 5 & 1 & 4 & 1 & ? \\ ? & 2 & 1 & 3 & ? & ? & 5 & 3 & 3 & 5 \\ 3 & 4 & ? & 2 & 5 & 5 & 3 & 2 & ? & 4 \\ 4 & 5 & 3 & 4 & 2 & 2 & 1 & ? & 5 & 5 \\ 2 & 4 & 2 & 5 & ? & 1 & 1 & 3 & 1 & 4 \\ ? & 1 & 4 & 4 & 3 & ? & 5 & 2 & 4 & 3 \end{bmatrix} \tag{50.138}$$

We set $M = 5$ (feature vectors h_i of size 5) and generate uniform random initial conditions for the variables $\{\boldsymbol{w}_{u,-1}, \boldsymbol{h}_{i,-1}, \boldsymbol{\theta}_u(-1), \boldsymbol{\alpha}_i(-1)\}$ in the open interval $(0,1)$.

We set $\rho = 0.001$. We normalize the entries of R to lie in the range $[0, 1]$ by replacing each numerical entry r by the value

$$r \leftarrow (r-1)/4 \tag{50.139}$$

where the denominator is the score range (highest value minus smallest value) and the numerator is subtracted from the smallest rating value (which is one). We repeat recursions (50.137) for 500 runs. At the end of the simulation, we use the parameters $\{w_u^\star, h_i^\star, \theta_u^\star, \alpha_i^\star\}$ to estimate each entry of R using

$$\widehat{r}_{ui} = (h_i^\star)^\mathsf{T} w_u^\star - \theta_u^\star - \alpha_i^\star \tag{50.140}$$

We undo the normalization by replacing each of these predicted values by

$$\widehat{r}_{ui} \leftarrow 4\,\widehat{r}_{ui} + 1 \tag{50.141}$$

and rounding $\widehat{r}_{ui}$ to the closest integer; scores above 5 are saturated at 5 and scores below 1 are fixed at 1. The result is the matrix $\widehat{R}$ shown below where we indicate the scores predicted for the unknown entries in red:

$$\widehat{R} = \begin{bmatrix} 4 & 3 & 3 & 3 & \mathbf{3} & 4 & 3 & \mathbf{3} & 3 & 3 \\ 4 & 4 & 3 & 3 & 3 & 4 & 2 & \mathbf{3} & 3 & \mathbf{3} \\ 3 & 3 & \mathbf{3} & 3 & 3 & 3 & 2 & 3 & 3 & 3 \\ \mathbf{4} & 4 & 3 & 3 & 3 & 4 & 2 & 3 & 3 & 3 \\ 3 & 3 & 2 & 2 & 2 & 3 & 1 & 2 & 2 & \mathbf{2} \\ \mathbf{2} & 2 & 3 & 3 & \mathbf{3} & \mathbf{2} & 4 & 3 & 3 & 3 \\ 4 & 4 & \mathbf{3} & 3 & 3 & 4 & 3 & 4 & \mathbf{3} & 3 \\ 4 & 4 & 3 & 3 & 3 & 4 & 3 & \mathbf{4} & 3 & 3 \\ 3 & 3 & 3 & 3 & \mathbf{3} & 3 & 3 & 3 & 3 & 3 \\ \mathbf{1} & 1 & 4 & 4 & 3 & \mathbf{1} & 5 & 2 & 4 & 3 \end{bmatrix} \tag{50.142}$$

Compared with the earlier result (16.61) obtained by applying a stochastic gradient procedure, we observe that the current simulation based on the alternating least-squares implementation is not able to recover several of the entries in the original matrix R. It is useful to recall that the risk function in (50.135) is not convex over the parameters and local minima are therefore possible. Figure 50.7 provides a color-coded representation of the entries of the original matrix R with the locations of the missing entries highlighted in red, and the recovered matrix $\widehat{R}$ on the right.

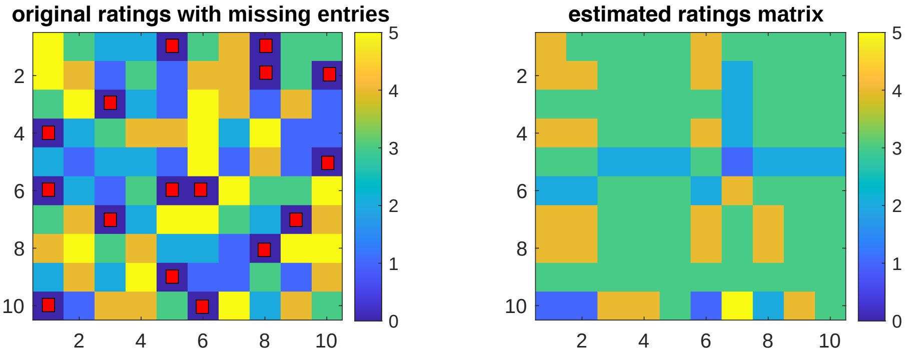

Figure 50.7 Color coded representation of the entries of the original matrix R with missing entries (*left*) and the recovered matrix $\widehat{R}$ (*right*).

We further denote the risk value at the start of each epoch of index k by

$$P(k) \triangleq \sum_{u=1}^{U} \rho\|w_u\|^2 + \sum_{i=1}^{I} \rho\|h_i\|^2 + \sum_{(u,i)\in\mathcal{R}} \Big(r_{ui} - h_i^{\mathsf{T}} w_u + \theta_u + \alpha_i\Big)^2 \tag{50.143}$$

where the parameters on the right-hand side are set to the values at the start of epoch k. Figure 50.8 plots the evolution of the risk curve (normalized by its maximum value so that its peak value is set to one).

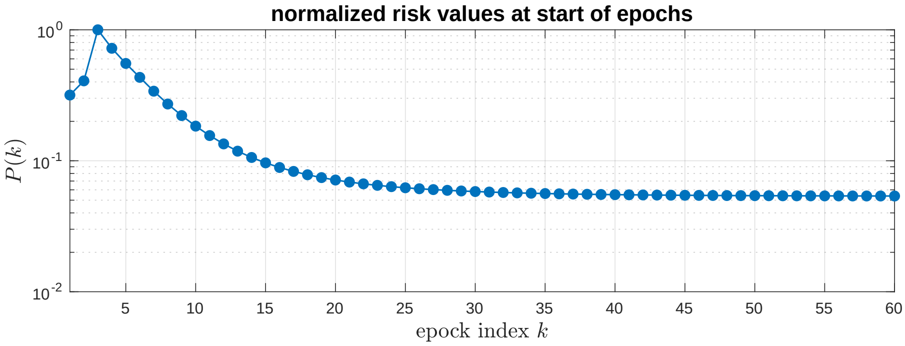

Figure 50.8 Evolution of the risk curve (50.143) with its peak value normalized to one.

50.4 IMPLICIT BIAS

We return to the standard least-squares problem (50.19), repeated here for ease of reference:

$$w^\star \triangleq \underset{w\in\mathbb{R}^M}{\text{argmin}} \; \|d - Hw\|^2 \tag{50.144}$$

and examine the case in which there are infinitely many solutions. In particular, we will assume $N < M$ so that H is a "fat" matrix with more columns than rows. This also means that there are fewer measurements than the size of w. We refer to this situation as the *under-determined* or *over-parameterized* least-squares problem.

It turns out that if we apply the traditional gradient-descent recursion to the solution of (50.144), namely,

$$\begin{aligned} w_n &= w_{n-1} - \mu\nabla_{w^{\mathsf{T}}} \|d - Hw\|^2 \Big|_{w=w_{n-1}} \\ &= w_{n-1} + 2\mu H^{\mathsf{T}}(d - Hw_{n-1}), \quad n \geq 0 \end{aligned} \tag{50.145}$$

where μ is a small step-size parameter, then the iterate w_n will converge to the minimum-norm solution, $w^\star = H^\dagger d$:

$$\lim_{n\to\infty} w_n = H^\dagger d \tag{50.146}$$

Proof of (50.146): Assume H has full row rank and introduce its singular value decomposition

$$H = U \begin{bmatrix} \Sigma & 0 \end{bmatrix} V^\mathsf{T}, \;\; UU^\mathsf{T} = I_N, \quad VV^\mathsf{T} = I_M \tag{50.147}$$

where Σ is $N \times N$ diagonal with positive singular values $\{\sigma_\ell^2 > 0\}$ for $\ell = 1, 2, \ldots, N$. We partition V into

$$V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}, \quad V_1 \in \mathbb{R}^{M\times N} \tag{50.148}$$

and note from the orthogonality of the $M \times M$ matrix V that

$$V^\mathsf{T} V = I_M \iff \begin{bmatrix} V_1^\mathsf{T} \\ V_2^\mathsf{T} \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} I_N & 0 \\ 0 & I_{M-N} \end{bmatrix} \tag{50.149}$$

Now, we know from result (50.179) that the minimum norm solution of the least-squares problem for the case under study is given by (recall (1.114)):

$$w^\star = H^\dagger d \; = \; V \begin{bmatrix} \Sigma^{-1} \\ 0 \end{bmatrix} U^\mathsf{T} d \tag{50.150}$$

We select the initial condition for the gradient-descent recursion (50.145) to lie in the range space of H^T, i.e.,

$$w_{-1} \in \mathcal{R}(H^\mathsf{T}) \iff w_{-1} = H^\mathsf{T} c, \;\; \text{for some } c \in \mathbb{R}^N \tag{50.151}$$

In this case, it is easy to see by iterating (50.145) that the successive w_n will remain in the range space of H^T:

$$w_n \in \mathcal{R}(H^\mathsf{T}), \quad n \geq 0 \tag{50.152}$$

Moreover, we can characterize the limit point of this sequence. For this purpose, we introduce a convenient change of variables in the form of the $M \times 1$ vector:

$$z_n \;\stackrel{\Delta}{=}\; V^\mathsf{T} w_n = \begin{bmatrix} V_1^\mathsf{T} w_n \\ V_2^\mathsf{T} w_n \end{bmatrix} \tag{50.153}$$

Multiplying recursion (50.145) by V^T from both sides leads to

$$z_n = z_{n-1} + 2\mu \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \Big(U^\mathsf{T} d - \begin{bmatrix} \Sigma & 0 \end{bmatrix} z_{n-1} \Big) \tag{50.154}$$

We partition z_n into $z_n = \text{col}\{a_n, b_n\}$ where the leading component a_n is $N \times N$. Then, the above relation gives:

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} = \begin{bmatrix} a_{n-1} \\ b_{n-1} \end{bmatrix} + 2\mu \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} (U^\mathsf{T} d - \Sigma a_{n-1}) \tag{50.155}$$

from which we conclude that

$$a_n = (I_N - 2\mu\Sigma^2) a_{n-1} + 2\mu \Sigma U^\mathsf{T} d \tag{50.156a}$$

$$b_n = b_{n-1} \tag{50.156b}$$

Observe that component b_n does not evolve with time and stays fixed at its initial value, denoted by

$$b_n \;\stackrel{\Delta}{=}\; b_2^\star = V_2^\mathsf{T} w_{-1} = V_2^\mathsf{T} H^\mathsf{T} c \stackrel{(50.149)}{=} 0 \tag{50.157}$$

On the other hand, the recursion for a_n has a diagonal coefficient matrix, $I_N - 2\mu\Sigma^2$. We can select μ to ensure this matrix is stable, namely, to guarantee

$$|1 - 2\mu\sigma_\ell^2| < 1, \;\; \forall\, \ell \iff \mu < 1/\sigma_{\max}^2 \tag{50.158}$$

in terms of the largest singular value of H. Under this condition, the recursion for a_n converges to the steady-state value

$$\lim_{n\to\infty} a_n \;\overset{\Delta}{=}\; a^\star \;=\; \Sigma^{-1}U^{\mathsf{T}}d \tag{50.159}$$

We therefore conclude that

$$\lim_{n\to\infty} z_n = \begin{bmatrix} \Sigma^{-1}U^{\mathsf{T}}d \\ 0 \end{bmatrix} \tag{50.160}$$

and, hence,

$$\lim_{n\to\infty} w_n = V\begin{bmatrix} \Sigma^{-1}U^{\mathsf{T}}d \\ 0 \end{bmatrix} \;=\; V\begin{bmatrix} \Sigma^{-1} \\ 0 \end{bmatrix} U^{\mathsf{T}}d \;=\; H^\dagger d \;\overset{(50.150)}{=}\; w^\star \tag{50.161}$$

as claimed.

∎

We therefore find that, in the under-determined case, when the amount of data available is smaller than the size of the parameter vector, the gradient-descent algorithm shows an *implicit bias* towards the minimum-norm solution. In other words, among all possible minimizers (and there are infinitely many in this case), the gradient-descent iteration converges to the minimum-norm solution. Other algorithms need not behave in the same manner and, therefore, the choice of the algorithm influences which parameter vector is ultimately learned.

50.5 COMMENTARIES AND DISCUSSION

Least-squares, Gauss, and RLS. The standard least-squares problem (50.19) has had an interesting and controversial history since its inception in the late 1700s, as already indicated in the texts by Kailath, Sayed, and Hassibi (2000) and Sayed (2003,2008). The criterion was formulated by the German mathematician **Carl Friedrich Gauss (1777–1855)** in 1795 at the age of 18 — see Gauss (1809). At that time, there was interest in a claim by the German philosopher **Georg Hegel (1770–1831)** who claimed that he has concluded using deductive logic that only seven planets existed. Then, on Jan. 1st, 1801, an astronomer noticed a moving object in the constellation of Aries, and the location of this celestial body was observed for 41 days before suddenly dropping out of sight. Gauss' contemporaries sought his help in predicting the future location of the heavenly body so that they could ascertain whether it was a planet or a comet (see Hall (1970), Plackett (1972), and Stigler (1981) for accounts of this story). With measurements available from the earlier sightings, Gauss formulated and solved a least-squares problem that could predict the location of the body (which turned out to be the planetoid Ceres). For some reason, Gauss did not bother to publish his least-squares solution, and controversy erupted in 1805 when the French mathematician **Adrien Legendre (1752–1833)** published a book where he independently invented the least-squares method — see Legendre (1805,1810). Since then, the controversy has been settled and credit is nowadays given to Gauss as the inventor of the method of least-squares. Interestingly, the method was also published around the same time by the Irish-American mathematician **Robert Adrain (1775–1843)** in the work by Adrain

(1808). Here is how Gauss himself motivated the least-squares problem:

> *"... if several quantities depending on the same unknown have been determined by inexact observations, we can recover the unknown either from one of the observations or from any of an infinite number of combinations of the observations. Although the value of an unknown determined in this way is always subject to error, there will be less error in some combinations than in others.... One of the most important problems in the application of mathematics to the natural sciences is to choose the best of these many combinations, i.e., the combination that yields values of the unknowns that are least subject to the errors."*
>
> *Extracted from Stewart (1995, pp. 31,33).*

Gauss' choice of the "best" combination was the one that minimizes the least-squares criterion. Actually, Gauss went further and formulated in his work on celestial bodies (ca. 1795) the unweighted ($\lambda = 1$) recursive-least-squares (RLS) solution, which we described in modern notation in (50.123). This step helped him save the trouble of having to solve a least-squares problem afresh every time a new measurement became available. Of course, Gauss' notation and derivation were reminiscent of the late 18th century mathematics and, therefore, they do not bear much resemblance with the linear algebraic and matrix arguments used in our derivation — see, e.g., the useful translation of Gauss' original work that appears in Stewart (1995). In modern times, the RLS algorithm is credited to Plackett (1950,1972). There is also an insightful and strong connection between RLS and Kalman filtering techniques, as detailed in Sayed and Kailath (1994) and in the textbooks by Sayed (2003,2008) — see Appendix 50.C further ahead.

Reliable numerical methods. There is a huge literature on least-squares problems and on reliable numerical methods for their solution — see, e.g., Higham (1996), Lawson and Hanson (1995), and Bjorck (1996). Among the most reliable methods for solving least-squares problems is the QR method, which is described in Prob. 50.5. The origin of the QR method goes back to Householder (1953, pp. 72–73), followed by Golub (1965), and Businger and Golub (1965). Since then, there has been an explosion of interest on solution methods for least-squares and recursive least-squares problems — see, for example, the treatment on array methods in Sayed (2003,2008).

LOWESS and LOESS smoothing. We described in Example 50.3 how localized least-squares formulations can be used to fit smooth curves onto data samples by means of the LOWESS and LOESS procedures, which were originally developed by Cleveland (1979) and Cleveland and Devlin (1988). These are simple but effective non-parametric techniques that slide a window over the data and fit locally either a regression line (LOWESS) or a quadratic curve (LOESS). The methods employ weighting to give more weight to data closer to the point that is being estimated and less weight to points that are farther away. We exhibited one choice for the weighting factor in (50.67) but other choices are possible, as explained in Cleveland (1979) where certain desirable properties on the weight factor are listed. These methods control the effect of outliers by re-scaling the weights and repeating the construction a few times.

Confidence intervals. We examined confidence intervals for least-squares problems in Example 50.4. In the derivation, we used (50.91) to conclude that the individual entries of the estimator $\boldsymbol{w}^\star$ are Gaussian and derived confidence intervals for them. If desired, we may alternatively work with the entire estimator $\boldsymbol{w}^\star$ (rather than its individual entries) and use expression (50.91) to describe an ellipsoidal region around $\boldsymbol{w}^\star$ where the true model is likely to lie with high confidence. It can be shown that for a significance level α (say, $\alpha = 5\%$), the true model w^o lies with $(1-\alpha)\%$ probability within the region

$$\text{ellipsoid} \;\stackrel{\Delta}{=}\; \left\{ w \;\middle|\; (w-w^\star)^\mathsf{T} H^\mathsf{T} H (w-w^\star) \le M\sigma_v^2\, F_\alpha^{(M,N-M)} \right\} \tag{50.162}$$

where the notation $F_\alpha^{(a,b)}$ refers to the point to the right of which the area under an $F-$distribution with parameters (a,b) is equal to α. This area is also called the *critical value* at which the significance level α is attained. For more discussion on confidence intervals and basic statistical concepts, the reader may refer to Draper and Smith (1998), Mendenhall, Beaver, and Beaver (2012), Witte and Witte (2013), and McClave and Sincich (2016).

Iterative reweighted least-squares. It is explained in Sayed (2003,2008) that the least-squares solution can also be useful in solving non-quadratic optimization problems of the form:

$$\min_{w\in\mathbb{R}^M} \left\{ \frac{1}{N}\sum_{n=0}^{N-1} |x(n)-y_n^\mathsf{T} w|^p \right\} \tag{50.163}$$

for some positive exponent p (usually $1 \le p \le 2$). This can be seen by reformulating the above criterion as a weighted least-squares problem in the following manner. Introduce the scalars (assumed nonzero):

$$r(n) \;\stackrel{\Delta}{=}\; |x(n)-y_n^\mathsf{T} w|^{p-2}, \qquad n=0,1,\ldots,N-1 \tag{50.164a}$$

and the diagonal weighting matrix

$$R = \text{diag}\Big\{ r(0), r(1), \ldots, r(N-1) \Big\} \tag{50.164b}$$

Then, the above optimization problem can be rewritten in the form

$$\min_{w\in\mathbb{R}^M} \; (d-Hw)^\mathsf{T} R (d-Hw) \tag{50.165}$$

where the vector d and the matrix H are defined as in (50.17). Of course, this reformulation is not truly a weighted least-squares problem because R is dependent on the unknown vector, w. Still, this rewriting of the risk function suggests the following iterative technique for seeking its minimizer. Given an estimate w_{k-1} at iteration $k-1$ we do the following:

$$\begin{aligned} \textbf{compute}\;\; r_k(n) = |x(n)-y_n^\mathsf{T} w_{k-1}|^{p-2}, &\qquad n=0,1,\ldots,N-1 \\ \textbf{set}\;\; R_k = \text{diag}\Big\{ r_k(0), r_k(1), \ldots, r_k(N-1) \Big\} & \\ \textbf{update}\text{ the estimate to }\; w_k = (H^\mathsf{T} R_k H)^{-1} H^\mathsf{T} R_k d & \\ \text{and }\textbf{repeat}\text{ until convergence} & \end{aligned} \tag{50.166}$$

This implementation assumes that the successive R_k are invertible. The algorithm is known as *iterative reweighted least-squares* (IRLS). It has several variations with improved stability and convergence properties (see, e.g., Osborne (1985) and Bjorck (1996). See also Fletcher, Grant, and Hebden (1971) and Kahng (1972)). One such variation is to evaluate w_k not directly as above but as a convex combination using the prior iterate w_{k-1} for some $0 < \beta \leq 1$ as follows:

$$\begin{aligned}
\textbf{compute}\ \ r_k(n) &= |x(n) - y_n^{\mathsf{T}} w_{k-1}|^{p-2}, \qquad n = 0, 1, \ldots, N-1 \\
\textbf{set}\ \ R_k &= \text{diag}\Big\{r_k(0), r_k(1), \ldots, r_k(N-1)\Big\} \\
\textbf{set}\ \ \overline{w}_k &= (H^{\mathsf{T}} R_k H)^{-1} H^{\mathsf{T}} R_k d \\
\textbf{set}\ \ w_k &= \beta \overline{w}_k + (1-\beta) w_{k-1} \\
&\text{and } \textbf{repeat} \text{ until convergence}
\end{aligned} \tag{50.167}$$

Matrix factorization. We described an *alternating* least-squares algorithm for the solution of the matrix factorization (or collaborative filtering) problem (50.135) in Example 50.6. We explained in the commentaries at the end of Chapter 16 that matrix factorization problems of this type arise in the design of recommender systems and were largely driven by the Netflix prize challenge, which ran during the period 2006-2009. Solution (50.137) is motivated by the works of Bell and Koren (2007a), Hu, Koren, and Volinsky (2008), Zhou *et al.* (2008), and Pilaszy, Zibriczky, and Tikk (2010). For more details on alternating methods, see also the treatment by Udell *et al.* (2016).

We recall that we encountered another instance of matrix factorization problems in the concluding remarks of Chapter 1 when we discussed the Eckart-Young theorem right after (1.222). The theorem dealt with the following scenario. Consider a $U \times I$ matrix R and assume we wish to determine a low-rank approximation for it in the form of the product $R \approx WH$, where W is $U \times M$, H is $M \times I$, and M is the desired rank approximation. The Eckart-Young theorem determines a collection of M column vectors $\{x_m, y_m\}$, where each x_m is $U \times 1$ and each y_m is $I \times 1$, in order to solve:

$$\widehat{R} \triangleq \underset{\{x_m, y_m\}}{\text{argmin}} \left\| R - \sum_{m=1}^{M} x_m y_m^{\mathsf{T}} \right\|_{\text{F}}^2 \tag{50.168}$$

Once the $\{x_m, y_m\}$ are determined, they are used to construct W and H as follows:

$$W = \begin{bmatrix} x_1 & x_2 & \ldots & x_M \end{bmatrix}, \quad H = \begin{bmatrix} y_1^{\mathsf{T}} \\ y_2^{\mathsf{T}} \\ \vdots \\ y_M^{\mathsf{T}} \end{bmatrix} \tag{50.169}$$

The solution of (50.168) requires all entries of R to be known (which obviously cannot be applied in the context of recommender systems where many entries are normally missing). The approximation $\widehat{R}$ is found as follows. We first introduce the SVD of R, say,

$$R = \sum_{n=1}^{r} \sigma_n u_n v_n^{\mathsf{T}} \tag{50.170}$$

where $r > M$ denotes the rank of R and the singular values $\{\sigma_n\}$ are ordered in decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$. Then, the solution to (50.168) is given by — recall Prob. 1.56:

$$\widehat{R} = \sum_{m=1}^{M} \sigma_m u_m v_m^{\mathsf{T}} \tag{50.171}$$

in terms of the singular vectors $\{u_m, v_m\}$ associated with the M largest singular values.

Sketching and randomized algorithms. We described in Example 50.5 some useful results on *randomized algorithms* and sketching applied to least-squares problems. These methods help to deal with situations involving massive amounts of data, while delivering some important performance guarantees. The basic idea, which involves projecting the data onto lower-dimensional spaces, is motivated by an important result from Johnson and Lindenstrauss (1984). In one of its simpler forms for Euclidean spaces, the result can be stated as follows.

Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss (1984)).*Consider a collection of M column vectors $\{x_m\}$ of dimension $N \times 1$ each. For any $0 < \epsilon < 1/2$, select a dimension $R = O((\log M)/\epsilon^2)$. Then, there exists a matrix $S \in \mathbb{R}^{R\times N}$ such that for all $m \neq m'$:*

$$1 - \epsilon \leq \frac{\|Sx_m - Sx_{m'}\|}{\|x_m - x_{m'}\|} \leq 1 + \epsilon \tag{50.172}$$

In the context of the least-squares problem studied in Example 50.5, the vectors x_m correspond to the columns of H or d. The above lemma essentially states that one can map a collection of vectors $\{x_n\}$ from an Euclidean space of high dimension N to another collection of vectors $\{Sx_m\}$ of much smaller dimension R such that the relative distance between any two points changes only by $1 \pm \epsilon$. This result has motivated a flurry of investigations on *sketching* methods. One notable advance was given by Sarlós (2006), who showed how to construct a sketching matrix S using fast Johnson-Lindenstrauss transforms leading to an ultimate complexity of $O(NM \log M)$ for the solution of least-squares problems. The Gaussian construction for a sketching matrix given in Example 50.5 is from Indyk and Motwani (1998), while the leverage-scores-based construction is from Drineas, Mahoney, and Muthukrishnan (2006b), and the Hadamard construction is from Ailon and Liberty (2013). Extensions to other convex problems appear in Pilanci and Wainwright (2015). Excellent surveys on randomized algorithms and sketching are given by Mahoney (2011) and Woodruff (2014) with derivations and justifications for several of the results and properties mentioned in the body of the chapter.

Implicit bias or regularization. We illustrated in Sec. 50.4 one instance of implicit bias (also called implicit regularization). We considered an over-parameterized least-squares problem where there are fewer data points than the size of the parameter vector, $w \in \mathbb{R}^M$. The analysis showed that the gradient-descent solution has an *implicit bias* towards the minimum-norm solution of the least-squares problem. Similar behavior occurs for other risk functions and is not limited to the least-squares case — see, e.g., Prob. 50.31 dealing with matrix factorization, the earlier Prob. 16.8 dealing with the Kaczmarz method, and future Prob. 61.7 dealing with logistic regression and support vector machines. Other algorithms need not behave in the same manner and may converge to other minimizers. Therefore, the choice of which algorithm to use has an influence on which model is learned in cases when a multiplicity of solutions exist. And some models are "better" than others because they may generalize better in the following sense. Once a solution $w^\star$ is found, the intent is to use it to predict target values x for future observations y that were not part of the original training data $\{d, H\}$ by using, for example, $\widehat{x}_t = y_t^\mathsf{T} w^\star$. The concept of "generalization" relates to how well a learned model $w^\star$ performs on new observations, i.e., how well it predicts. We will discuss generalization in the context of classification problems in greater detail in future Chapter 64. For more discussion on the topic of implicit bias in the machine learning literature, the reader may refer to Gower and Richtárik (2015), Neyshabur, Tomioka, and Srebro (2015), Gunasekar *et al.* (2017,2018), Soudry *et al.* (2018), Jin and Montúfar (2020), and the references therein.

Recursive least-squares and Kalman filtering. Following Sayed and Kailath (1994) and Sayed (2003,2008), Appendix 50.B describes a useful equivalence result between stochastic and deterministic estimation problems with quadratic risks. The equivalence is then used in Appendix 50.C, based on arguments from Kailath, Sayed, and Hassibi (2000) and Sayed (2003,2008), to clarify the fundamental connection that exists between recursive least-squares and Kalman filtering, so much so that solving a problem in one domain is equivalent to solving a problem in the other domain. One of the earliest mentions of a relation between least-squares and Kalman filtering appears to be Ho (1963); however, this reference considers only a special estimation problem where the successive observation vectors are identical. Later references are Sorenson (1966) and Aström and Wittenmark (1971); these works focus only on the standard (i.e., unregularized) least-squares problem, in which case an exact relationship between least-squares and Kalman filtering does not actually exist, especially during the initial stages of adaptation when the least-squares problem is under-determined. Soon afterwards, in work on channel equalization, Godard (1974) rephrased the growing-memory (i.e., $\lambda = 1$) RLS problem in a stochastic state-space framework, with the unknown state corresponding to the unknown weight vector in a manner similar to what we encountered in Example 30.4. Similar constructions also appeared in Willsky (1979), Anderson and Moore (1979), Ljung (1987), Strobach (1990), and Söderström (1994). In the works by Anderson and Moore (1979), Ljung (1987), and Söderström (1994), the underlying models went a step further and incorporated the case of exponentially decaying memory (i.e., $\lambda < 1$) by formulating state-space models with a time-variant noise variance. Nevertheless, annoying discrepancies persisted that precluded a direct correspondence between the exponentially-weighted RLS ($\lambda < 1$) and the Kalman variables. Some of these discrepancies were overcome essentially by fiat (see, e.g., the treatment by Haykin (1991)). This lack of a direct correspondence may have inhibited application of the extensive body of Kalman filter results to the adaptive least-squares problem until a resolution was given in the work by Sayed and Kailath (1994). In retrospect, by a simple device, the latter reference was able to obtain a perfectly matched state-space model for the case of exponentially decaying memory ($\lambda < 1$), with a direct correspondence between the variables in the exponentially weighted RLS problem and the variables in the state-space estimation problem.

Sea-level and global temperature changes. Figure 50.2 illustrates the result of fitting a linear regression model onto measurements of sea level changes. The source of the data is the NASA Goddard Space Flight Center at https://climate.nasa.gov/vital-signs/sea-level/. For more information on how the data was generated, the reader may consult Beckley *et al.* (2017) and the report GSFC (2017). Similarly, Fig. 50.3 illustrates the fitting of LOWESS and LOESS smoothing curves onto measurements of changes in the global surface temperature. The source of the data is the NASA Goddard Institute for Space Studies (GISS) at https://climate.nasa.gov/vital-signs/global-temperature/.

PROBLEMS[1]

50.1 Consider an $N \times M$ full-rank matrix H with $N \geq M$, and two column vectors d and z of dimensions $N \times 1$ each. Let $\widetilde{d} = \mathcal{P}_H^{\perp} d$ and $\widetilde{z} = \mathcal{P}_H^{\perp} z$. Are the residual vectors $\widetilde{d}$ and $\widetilde{z}$ collinear in general? If your answer is positive, justify it. If the answer is negative, can you give conditions on N and M under which $\widetilde{d}$ and $\widetilde{z}$ will be collinear?

50.2 Let H be $N \times M$ with full-column rank. Show that any vector in the column span of $\mathcal{P}_H^{\perp}$ is orthogonal to any vector in the column span of H. That is, show that $H^{\mathsf{T}} \mathcal{P}_H^{\perp} = 0$.

50.3 Consider the standard least-squares problem (50.19). Comment on the solution $w^{\star}$ in the following three cases: (a) $d \in \mathcal{N}(H)$, (b) $d \in \mathcal{R}(H)$, and (c) $d \in \mathcal{N}(H^{\mathsf{T}})$.

[1] Several problems in this section are adapted exercises from Sayed (2003,2008).

50.4 Solving the normal equations $H^{\mathsf{T}}Hw^\star = H^{\mathsf{T}}d$ by forming the matrix $H^{\mathsf{T}}H$ (i.e., by squaring the data) is a bad idea in general. Consider the full-rank matrix

$$H = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \\ 1 & 1 \end{bmatrix}$$

where ϵ is a very small positive number that is of the same order of magnitude as machine precision. Assuming $2+\epsilon^2 = 2$ in finite precision, what is the rank of $H^{\mathsf{T}}H$?

50.5 A numerically-reliable method for solving the normal equations $H^{\mathsf{T}}Hw^\star = H^{\mathsf{T}}d$ is the QR method. It avoids forming the product $H^{\mathsf{T}}H$, which is problematic for ill-conditioned matrices. The QR method works directly with H and uses its QR decomposition — defined earlier in Sec. 1.6:

$$H = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$$

where Q is $N \times N$ orthogonal and R is $M \times M$ upper-triangular with positive diagonal entries. Let $\text{col}\{z_1, z_2\} = Q^{\mathsf{T}}d$, where z_1 is $M \times 1$. Verify that

$$\|d - Hw\|^2 = \|z_1 - Rw\|^2 + \|z_2\|^2$$

Refer to the standard least-squares problem (50.19) and verify that the least-squares solution $w^\star$ can be obtained by solving the triangular linear system of equations $Rw^\star = z_1$. Conclude that the minimum risk is $\|z_2\|^2$.

50.6 Refer to Example 50.1 but assume now that the zero-mean Gaussian noise process is colored. Collect the noise terms into the column vector

$$\boldsymbol{v} \;\stackrel{\Delta}{=}\; \text{col}\Big\{\boldsymbol{v}(0), \boldsymbol{v}(1), \ldots, \boldsymbol{v}(N-1)\Big\}$$

and denote its covariance matrix by $R_v = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} > 0$. Use the data and vector notation (50.17) to verify that the maximum-likelihood estimate for w is the solution to the weighted least-squares problem:

$$\min_{w\in\mathbb{R}^M} \|d - Hw\|^2_{R_v^{-1}} \quad\Longrightarrow\quad w^\star = (H^{\mathsf{T}}R_v^{-1}H)^{-1}H^{\mathsf{T}}R_v^{-1}d$$

where the notation $\|a\|^2_R$ stands for $a^{\mathsf{T}}Ra$.

50.7 Let $\xi(n)$ denote the minimum risk value of (50.134) with w replaced by w_n.

(a) Show that $\xi(n) = d_n^{\mathsf{T}}\Lambda_n(d_n - H_n w_n)$.

(b) Derive the time-update relation $\xi(n) = \lambda\xi(n-1) + t(n)e^2(n)$, $\xi(-1) = 0$.

50.8 Consider an ℓ_2-regularized least-squares problem of the form:

$$\underset{w\in\mathbb{R}^M,\theta\in\mathbb{R}}{\text{argmin}} \left\{\rho\|w\|^2 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}}w + \theta\Big)^2\right\}$$

Observe that regularization is applied to w only and not to θ. Introduce the sample averages:

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x(n), \qquad \bar{y} = \frac{1}{N}\sum_{n=0}^{N-1} y_n$$

(a) Fix w and show that optimizing over θ leads to the expression $\theta = \bar{y}^{\mathsf{T}}w - \bar{x}$.

(b) Center the data and define $x'(n) = x(n) - \bar{x}$ and $y_n' = y_n - \bar{y}$. Conclude that the above least-squares problem is equivalent to solving a traditional regularized problem without offset, namely,

$$\underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{\rho\|w\|^2 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x'(n) - (y_n')^{\mathsf{T}}w\Big)^2\right\}$$

50.9 Let $w^\star$ and $w^\star_{\rm reg}$ denote the solutions to the following problems:

$$\begin{aligned} w^\star &\triangleq \underset{w\in{\rm I\!R}^M}{\text{argmin}} \;\; \|d-Hw\|^2 \\ w^\star_{\rm reg} &\triangleq \underset{w\in{\rm I\!R}^M}{\text{argmin}} \;\; \Big\{\rho\|w\|^2+\|d-Hw\|^2\Big\}, \quad \rho>0 \end{aligned}$$

Let $Q=H^\mathsf{T}H$, assumed invertible. Show that $w^\star_{\rm reg}=(I_M+\rho Q)^{-1}w^\star$.

50.10 Consider the weighted least-squares problem (50.49). Verify that the orthogonality condition in this case is given by

$$H^\mathsf{T}R(d-Hw^\star)=0 \iff H^\mathsf{T}R\widetilde{d}=0$$

where $\widetilde{d}=d-\widehat{d}$ and $\widehat{d}=Hw^\star$. Show further that the minimum risk is given by $\xi=d^\mathsf{T}R\widetilde{d}$.

50.11 Refer to the stochastic model (50.88) where $\boldsymbol{v}$ has covariance matrix $\sigma_v^2 I_N$ but is not necessarily Gaussian. Relation (50.89) will continue to hold, linking the true model w^o to the least-squares model $w^\star$. Introduce the mean-square error risk, $P(w)=\mathbb{E}\,\|\boldsymbol{d}-Hw\|^2$, where the expectation is over the source of randomness in $\boldsymbol{d}$.

(a) Let $\widetilde{w}=w^o-w$. Verify that $P(w)=\widetilde{w}^\mathsf{T}H^\mathsf{T}H\widetilde{w}+N\sigma_v^2$. Conclude that the minimum value is attained at $w=w^o$ and is equal to $P(w^o)=N\sigma_v^2$.

(b) Using expression (50.89) verify that the least-squares solution $\boldsymbol{w}^\star$, which is now random since it depends on $\boldsymbol{d}$, leads to an average excess risk value of $\mathbb{E}\,P(\boldsymbol{w}^\star)-P(w^o)=\sigma_v^2 M$, which is dependent on the problem dimension, M.

50.12 We continue with the stochastic model (50.88), but assume now that the rows of H are Gaussian distributed with zero mean and unit covariance matrix, i.e., each $\boldsymbol{y}_n\sim\mathcal{N}_{\boldsymbol{y}_n}(0,I_M)$. We continue to assume that $\boldsymbol{v}$ has zero mean and covariance matrix $\sigma_v^2 I_N$ and is independent of $\boldsymbol{H}$. In this problem we consider both situations in which $N\geq M$ (overdetermined least-squares, with more data than unknowns) and $N<M$ (under-determined or over-parameterized least-squares). Introduce the weight-error vector $\widetilde{\boldsymbol{w}}=w^o-\boldsymbol{w}^\star$, where $\boldsymbol{w}^\star$ is a least-squares solution.

(a) Assume first that $N\geq M$ and show that

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}\|^2=\sigma_v^2\,\mathbb{E}\Big\{\text{Tr}(\boldsymbol{H}^\mathsf{T}\boldsymbol{H})\Big\} \;=\; \frac{\sigma_v^2 M}{N-M-1}, \quad \text{for } N\geq M+2$$

where we are denoting $\boldsymbol{H}$ in boldface since its entries are now random.

(b) Assume next that $N<M$ and let $\boldsymbol{w}^\star$ refer to the minimum-norm least-squares solution. Show that

$$\begin{aligned} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}\|^2 &= \mathbb{E}\,\|(I_M-\boldsymbol{H}^\mathsf{T}(\boldsymbol{H}\boldsymbol{H}^\mathsf{T})^{-1}\boldsymbol{H})w^o\|^2+\sigma_v^2\,\mathbb{E}\Big\{\text{Tr}(\boldsymbol{H}\boldsymbol{H}^\mathsf{T})\Big\} \\ &= \frac{M-N}{M}\|w^o\|^2 \;+\; \frac{\sigma_v^2 N}{M-N-1}, \quad \text{for } M\geq N+2 \end{aligned}$$

(c) Compare both situations as M varies.

Remark. The result of this problem, and especially the result in part (c) showing how the mean-square error behaves as a function of increasing complexity M, is related to the phenomena of *double descent* and bias-variance tradeoff in learning — see, e.g., Belkin, Sa, and Mandal (2018), Belkin, Rakhlin, and Tsybakov (2019), Hastie *et al.* (2019), and Mei and Montanari (2020). To solve the problem, the reader needs to rely on some properties of the Wishart distribution. Consider a collection of $M-$dimensional vectors $\{\boldsymbol{a}_n\}$, each arising from a zero-mean Gaussian distribution with covariance matrix $\Sigma>0$, i.e., $\boldsymbol{a}\sim\mathcal{N}_{\boldsymbol{a}}(0,\Sigma)$. Let $\boldsymbol{X}=\sum_{n=1}^N \boldsymbol{a}_n\boldsymbol{a}_n^\mathsf{T}$, which is $M\times M$. Then, for $N\geq M$, it is known that $\boldsymbol{X}$ is invertible almost surely and it follows a so-called *Wishart distribution* with mean zero, N degrees of freedom, and scale parameter Σ, written as $\boldsymbol{X}\sim\mathcal{W}(N,\Sigma)$. Its mean is $\mathbb{E}\,\boldsymbol{X}=N\Sigma$. The inverse matrix $\boldsymbol{X}^{-1}$ follows an inverse

Wishart distribution with mean zero, N degrees of freedom, and scalar parameter Σ^{-1}, written as $\boldsymbol{X}^{-1} \sim \mathcal{W}^{-1}(N, \Sigma^{-1})$. The respective pdfs are proportional to

$$f_{\boldsymbol{X}}(X) \propto \Big(\det \boldsymbol{X}\Big)^{(N-M-1)/2} \times \exp\Big\{-\frac{1}{2}\text{Tr}(\Sigma^{-1}X)\Big\}, \quad N \geq M$$
$$f_{\boldsymbol{X}^{-1}}(X^{-1}) \propto \Big(\det \boldsymbol{X}^{-1}\Big)^{-(N+M+1)/2} \times \exp\Big\{-\frac{1}{2}\text{Tr}(\Sigma^{-1}X^{-1})\Big\}, \quad N \geq M$$

For more information on the Wishart distribution, the reader may refer to Eaton (1983), Gupta and Nagar (2000), and Anderson (2003).

50.13 Refer to the stochastic model (50.88) where $\boldsymbol{v}$ has covariance matrix $\sigma_v^2 I_N$. Show that

$$\mathbb{E}\,\|H(\boldsymbol{w}^\star - w^o)\|^2 \;\leq\; 4\sigma_v^2\,\text{rank}(H)$$

Remark. See Rigollet and Huetter (2017) for a related discussion.

50.14 Consider a symmetric positive-definite weighting matrix, R, and a symmetric positive-definite regularization matrix, Π. Verify that the "normal equations" that describe all solutions to the regularized and weighted least-squares problem:

$$\min_{w\in\mathbb{R}^M} \Big\{w^\mathsf{T}\Pi w \;+\; (d-Hw)^\mathsf{T}R(d-Hw)\Big\}$$

are given by $(\Pi + H^\mathsf{T}RH)w^\star = H^\mathsf{T}Rd$. Verify that the "orthogonality condition" in this case amounts to requiring:

$$H^\mathsf{T}R(d - Hw^\star) = \Pi w^\star \iff H^\mathsf{T}R\widetilde{d} = \Pi w^\star$$

where $\widetilde{d} = d - \widehat{d}$ and $\widehat{d} = Hw^\star$. Show further that the minimum cost is given by either expression:

$$\xi = d^\mathsf{T}R\widetilde{d} = d^\mathsf{T}(R^{-1} + H\Pi^{-1}H^\mathsf{T})^{-1}d$$

50.15 In constrained least-squares problems we seek to minimize $\|d - Hw\|^2$ over $w \in \mathbb{R}^M$ subject to the linear constraint $Aw = b$, where the data matrices H and A have dimensions $N \times M$ ($N \geq M$) and $P \times M$ ($P \leq M$), respectively. Both matrices $\{H, A\}$ are assumed to have full rank. Note that H is "tall" while A is "fat." Show that the solution is given by

$$w_c^\star = w^\star - (H^\mathsf{T}H)^{-1}A^\mathsf{T}\Big(A(H^\mathsf{T}H)^{-1}A^\mathsf{T}\Big)^{-1}(Aw^\star - b)$$

where $w^\star$ is the standard least-squares solution, $w^\star = (H^\mathsf{T}H)^{-1}H^\mathsf{T}d$.

50.16 Consider a data matrix H and partition it as $H = [d \;\; \bar{H} \;\; z]$, with d and z denoting its leading and trailing columns, respectively. Let $\widehat{d}$ and $\widehat{z}$ denote the regularized least-squares estimates of d and z given $\bar{H}$, namely, $\widehat{d} = \bar{H}w_y^\star$, $\widehat{z} = \bar{H}w_z^\star$, $\widetilde{d} = d - \widehat{d}$, and $\widetilde{z} = z - \widehat{z}$, where $w_y^\star$ and $w_z^\star$ are the solutions of

$$\min_{w_y} \Big\{w_y^\mathsf{T}\Pi w_y + \|d - \bar{H}w_y\|^2\Big\} \quad\text{and}\quad \min_{w_z} \Big\{w_z^\mathsf{T}\Pi w_z + \|z - \bar{H}w_z\|^2\Big\}$$

for some positive-definite matrix Π. Show that $(\widetilde{d})^\mathsf{T}z = d^\mathsf{T}\widetilde{z}$. Define $\kappa \overset{\Delta}{=} (\widetilde{d})^\mathsf{T}\widetilde{z}/(\|\widetilde{d}\|\;\|\widetilde{z}\|)$. Show that $|\kappa| \leq 1$.

50.17 Refer to the recursive least-squares (RLS) algorithm in Sec. 50.3. Introduce the *a-priori* and *a-posteriori* errors $e(n) = x(n) - y_n^\mathsf{T}w_{n-1}$ and $r(n) = x(n) - y_n^\mathsf{T}w_n$. Observe that one error depends on w_{n-1} while the other error depends on the updated iterate, w_n. The conversion factor allows us to transform $e(n)$ into $r(n)$ without the need to update w_{n-1} to w_n. Show that $r(n) = t(n)e(n)$. Conclude that $|r(n)| \leq |e(n)|$.

50.18 Refer to the derivation of the exponentially-weighted least-squares algorithm in Sec. 50.3 but assume now that d_N evolves in time in the following manner:

$$d_N = \begin{bmatrix} ad_{N-1} \\ x(N) \end{bmatrix}$$

for some scalar a. The choice $a = 1$ reduces to the situation studied in the body of the chapter. Show that the solution w_N, and the corresponding minimum cost, $\xi(N)$, can be computed recursively as follows. Start with $w_{-1} = 0$, $P_{-1} = (1/\rho')I$, and $\xi(-1) = 0$, and iterate for $n \geq 0$:

$$\begin{aligned}
t(n) &= 1/(1 + \lambda^{-1} y_n^{\mathsf{T}} P_{n-1} y_n) \\
g_n &= \lambda^{-1} P_{n-1} y_n t(n) \\
e(n) &= x(n) - a y_n^{\mathsf{T}} w_{n-1} \\
w_n &= a w_{n-1} + g_n e(n) \\
P_n &= \lambda^{-1} P_{n-1} - g_n g_n^{\mathsf{T}} / t(n) \\
\xi(n) &= \lambda a^2 \xi(n-1) \; + \; t(n) e^2(n)
\end{aligned}$$

In particular, observe that the scalar a appears in the expressions for $\{w_n, e(n), \xi(n)\}$. Show further that $r(n) = t(n)e(n)$ where $r(n) = x(n) - y_n^{\mathsf{T}} w_n$.

50.19 All variables are scalars. Consider N noisy measurements of an unknown x, say, $d(n) = x + v(n)$, and formulate the following two optimization problems:

$$\widehat{x}_{\text{mean}} \;\stackrel{\Delta}{=}\; \underset{x}{\text{argmin}} \;\; \frac{1}{N} \sum_{n=1}^{N} (d(n) - x)^2, \quad \widehat{x}_{\text{median}} \;\stackrel{\Delta}{=}\; \underset{x}{\text{argmin}} \;\; \frac{1}{N} \sum_{n=1}^{N} |d(n) - x|$$

(a) Show that $\widehat{x}_{\text{mean}}$ is the sample mean, i.e., $\widehat{x}_{\text{mean}} = \frac{1}{N} \sum_{n=1}^{N} d(n)$.

(b) Show that $\widehat{x}_{\text{median}}$ is the median of the observations, where the median is such that an equal number of observations exists to its left and to its right.

50.20 At each time $n \geq 0$, M noisy measurements of a scalar unknown variable x are collected from M spatially-distributed sensors, say, $d_m(n) = x + v_m(n), m = 0, 1, \ldots, M-1$. The unknown x is estimated by solving a least-squares problem of the form:

$$\widehat{x}_N \;\stackrel{\Delta}{=}\; \underset{x}{\text{argmin}} \left\{ \sum_{n=0}^{N} \lambda^{N-n} \left(\sum_{m=0}^{M-1} \alpha_m(n) \, |d_m(n) - x|^2 \right) \right\}$$

where $0 \ll \lambda \leq 1$ is an exponential forgetting factor and the $\{\alpha_k(n)\}$ are some nonnegative weighting coefficients. Show that $\widehat{x}_N$ can be computed recursively as follows:

$$\begin{aligned}
\phi(n) &= \lambda \phi(n-1) + \sum_{m=0}^{M-1} \alpha_m(n), \quad \phi(-1) = 0 \\
s(n) &= \lambda s(n-1) + \sum_{m=0}^{M-1} \alpha_m(n) d_m(n), \quad s(-1) = 0 \\
\widehat{x}_n &= s(n)/\phi(n)
\end{aligned}$$

50.21 Two least-squares estimators are out of sync. At any time N, estimator #1 computes the estimate $w_{1,0:N-1}$ that corresponds to the solution of

$$w_{1,0:N-1} \;\stackrel{\Delta}{=}\; \underset{w \in \mathbb{R}^M}{\text{argmin}} \;\; \left\{ \rho' \lambda^N \|w\|^2 \; + \; \sum_{n=0}^{N-1} \lambda^{N-1-n} (x(n) - y_n^{\mathsf{T}} w)^2 \right\}$$

where $\rho' > 0$ and λ is the forgetting factor. Note that $w_{1,0:N-1}$ is an estimate that is based on measurements between times $n = 0$ and $n = N - 1$. On the other hand, estimator #2 computes the estimate $w_{2,1:N}$ that corresponds to the solution of

$$w_{2,1:N} \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\lambda^{N}\|w\|^2 \;+\; \sum_{n=1}^{N} \lambda^{N-n}(x(n) - y_n^{\mathsf{T}} w)^2 \right\}$$

Here, $w_{2,1:N}$ is an estimate that is based on measurements between times $n = 1$ and $n = N$. Can you use the available estimates $\{w_{1,0:N-1}, w_{2,1:N}, N \geq 0\}$ to construct the recursive solution of

$$w_N \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\lambda^{N+1}\|w\|^2 \;+\; \sum_{n=0}^{N} \lambda^{N-n}(x(n) - y_n^{\mathsf{T}} w)^2 \right\}$$

where w_N is an estimate that is based on all data up to time N? If so, explain the construction. If not, explain why not.

50.22 Node #1 observes even-indexed data $\{x(2n), y_{2n}\}$ for $n \geq 0$ and computes the recursive least-squares solution of

$$w_{2n} \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\,\lambda^{2n+1}\,\|w\|^2 \;+\; \sum_{j=0}^{n} \lambda^{2n-2j}(x(2j) - y_{2j}^{\mathsf{T}} w)^2 \right\}$$

where $\rho' > 0$ is a regularization factor and λ is the forgetting factor. Note that w_{2n} is an estimate that is based solely on the even-indexed data. Likewise, node #2 observes odd-indexed data $\{x(2n+1), y_{2n+1}\}$ for $n \geq 0$ and computes the recursive least-squares solution of

$$w_{2n+1} \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\,\lambda^{2n+2}\,\|w\|^2 \;+\; \sum_{j=0}^{n} \lambda^{2n-2j}(x(2j+1) - y_{2j+1}^{\mathsf{T}} w)^2 \right\}$$

Here, w_{2n+1} is an estimate that is based solely on the odd-indexed data. Can you use the available estimates $\{w_{2n}, w_{2n+1}, n \geq 0\}$ to construct the recursive solution of

$$w_N \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\,\lambda^{N+1}\,\|w\|^2 \;+\; \sum_{j=0}^{N} \lambda^{N-j}(x(j) - y_j^{\mathsf{T}} w)^2 \right\}$$

where w_N is an estimate that is based on all data (both even and odd-indexed) up to time N? If so, explain the construction. If not, explain why not.

50.23 Consider the optimization problem

$$w_N \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{ \rho'\lambda^{N+1}\|w\|^2 + \mathbb{E}\left(\sum_{n=0}^{N} \lambda^{N-n}(x(n) - \boldsymbol{\alpha}\, y_n^{\mathsf{T}} w)^2 \right) \right\}$$

where the data $\{x(n), y_n\}$ are deterministic measurements with $x(n)$ a scalar and y_n a column vector of size $M \times 1$. The random variable $\boldsymbol{\alpha}$ is Bernoulli and assumes the value $\boldsymbol{\alpha} = 1$ with probability p and the value $\boldsymbol{\alpha} = 0$ with probability $1 - p$; it is used to model a faulty sensor – when the sensor fails, no data is measured. Let w_N denote the solution. Can you determine a recursion to go from w_{N-1} to w_N?

50.24 Consider an unknown $M \times 1$ vector $w = \text{col}\{w_1, w_2\}$, where w_1 is $L \times 1$. Introduce the least-squares problem:

$$\min_{w\in\mathbb{R}^M} \left\{ w_1^{\mathsf{T}} \Pi w_1 + \|z_N - H_N w\|^2 + \|d_N - G_N w_1\|^2 \right\}$$

where $\Pi > 0$,

$$z_N = \begin{bmatrix} z(0) \\ z(1) \\ \vdots \\ z(N) \end{bmatrix}, \quad d_N = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N) \end{bmatrix}, \quad H_N = \begin{bmatrix} y_0^\mathsf{T} \\ y_1^\mathsf{T} \\ \vdots \\ y_N^\mathsf{T} \end{bmatrix}, \quad G_N = \begin{bmatrix} s_0^\mathsf{T} \\ s_1^\mathsf{T} \\ \vdots \\ s_N^\mathsf{T} \end{bmatrix}$$

Let w_N denote the solution and let $\xi(N)$ be the resulting minimum cost.
(a) Relate w_N to w_{N-1}.
(b) Relate $\xi(N)$ to $\xi(N-1)$.

50.25 Let $w^\star$ denote the solution to the following regularized least-squares problem

$$\min_{w\in{\rm I\!R}^M} \left\{ w^\mathsf{T}\Pi w \;+\; (d-Hw)^\mathsf{T}R(d-Hw) \right\}$$

where $R > 0$ and $\Pi > 0$. Let $\widehat{d} = Hw^\star$ denote the resulting estimate of d and let ξ denote the corresponding minimum cost. Now consider the extended problem

$$\min_{w_z\in{\rm I\!R}^{M+1}} \left\{ w_z^\mathsf{T}\Pi_z w_z \;+\; \left\| \begin{bmatrix} d \\ \gamma \end{bmatrix} - \begin{bmatrix} h_a & H & h_b \\ \alpha_a & h^\mathsf{T} & \alpha_b \end{bmatrix} w_z \right\|^2_{R_z} \right\}$$

where $\{h, h_a, h_b\}$ are column vectors, $\{\gamma, \alpha_a, \alpha_b, a, b\}$ are scalars, and

$$\Pi_z = \begin{bmatrix} a & & \\ & \Pi & \\ & & b \end{bmatrix}, \quad R_z = \begin{bmatrix} R & \\ & 1 \end{bmatrix}$$

Let

$$\widehat{d}_z = \begin{bmatrix} h_a & H & h_b \\ \alpha_a & h^\mathsf{T} & \alpha_b \end{bmatrix} w_z^\star$$

and let ξ_z denote the corresponding minimum risk of the extended problem. Relate $\{w_z^\star, \widehat{d}_z, \xi_z\}$ to $\{w^\star, \widehat{d}, \xi\}$.

50.26 Consider an $M \times m$ full-rank matrix A ($M > m$) and let w be any vector in its range space, i.e., $w \in \mathcal{R}(A)$. Let w_N denote the solution to the following regularized least-squares problem:

$$\min_{w\in\mathcal{R}(A)} \left\{ \lambda^{N+1} w^\mathsf{T}\Pi w + \sum_{n=0}^{N} \lambda^{N-n}(x(n) - y_n^\mathsf{T}w)^2 \right\}$$

where $\Pi > 0$ and y_n is $M \times 1$. Find a recursion relating w_N to w_{N-1}.

50.27 Consider a least-squares problem of the form

$$\min_{w\in{\rm I\!R}^M} \left\{ \rho\|w\|^2 \;+\; \sum_{n=0}^{N} \lambda^{N-n}|x(n) - y_n^\mathsf{T}w|^2 \right\}$$

where $\rho > 0$ is a regularization parameter, y_n is an $M \times 1$ regression vector, and $0 \ll \lambda \le 1$ is a forgetting factor defined as follows:

$$\lambda = \begin{cases} \lambda_e, & \text{for } n \text{ even} \\ \lambda_o, & \text{for } n \text{ odd} \end{cases}$$

Let w_N denote the solution to the above least-squares problem. Derive a recursive solution that updates w_N to w_{N+1}?

50.28 Consider a regularized least-squares problem of the form

$$\min_{w\in\mathbb{R}^M} \Big\{(w-\bar{w})^\mathsf{T}\Pi(w-\bar{w}) \;+\; (z_{B-1}-\mathcal{H}_{B-1}w)^\mathsf{T}\mathcal{R}_{B-1}(z_{B-1}-\mathcal{H}_{B-1}w)\Big\}$$

where $\Pi > 0$, $\mathcal{R}_{B-1} > 0$ is a weighting matrix, and $\bar{w}$ is some known initial condition. We partition the entries of $\{z_{B-1}, \mathcal{H}_{B-1}\}$ into block vectors and block matrices:

$$z_{B-1} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{B-1} \end{bmatrix}, \qquad \mathcal{H}_{B-1} = \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{B-1} \end{bmatrix}$$

where each d_b has dimensions $p \times 1$ and each U_b has dimensions $p \times M$. We further assume that the positive-definite weighting matrix $\mathcal{R}_{B-1}$ has a block diagonal structure, with $p \times p$ positive-definite diagonal blocks, say $\mathcal{R}_{B-1} = \text{blkdiag}\{R_0^{-1}, R_1^{-1}, \ldots, R_{B-1}^{-1}\}$. Let w_{B-1} denote the solution of the above least-squares problem and let $P_{B-1} = (\Pi + \mathcal{H}_{B-1}^\mathsf{T}\mathcal{R}_{B-1}\mathcal{H}_{B-1})^{-1}$.

(a) Show that $P_B = P_{B-1} - P_{B-1}U_B^\mathsf{T}T_BU_BP_{B-1}$, with initial condition $P_{-1} = \Pi^{-1}$ and where $T_B = (R_B + U_BP_{B-1}U_B^\mathsf{T})^{-1}$.

(b) Show that $w_B = w_{B-1} + P_{B-1}U_B^\mathsf{T}T_B(d_B - U_Bw_{B-1})$.

(c) Conclude that w_B can be computed recursively by means of the following block RLS algorithm. Start with $w_{-1} = \bar{w}$ and $P_{-1} = \Pi^{-1}$ and repeat for $b \geq 0$:

$$\left\{\begin{array}{lcl} T_b & = & (R_b + U_bP_{b-1}U_b^\mathsf{T})^{-1} \\ G_b & = & P_{b-1}U_b^\mathsf{T}T_b \\ w_b & = & w_{b-1} + G_b(d_b - U_bw_{b-1}) \\ P_b & = & P_{b-1} - G_bT_b^{-1}G_b^\mathsf{T} \end{array}\right.$$

(d) Establish the equalities $G_B = P_BU_B^\mathsf{T}R_B^{-1}$ and $T_B = R_B^{-1} - R_B^{-1}U_BP_BU_B^\mathsf{T}R_B^{-1}$.

(e) Let $\{r_B, e_B\}$ denote the *a-posteriori* and *a-priori* error vectors, $r_B = d_B - U_Bw_B$ and $e_B = d_B - U_Bw_{B-1}$. Show that $R_B^{-1}r_B = T_Be_B$.

(f) Let $\xi(B-1)$ denote the minimum cost associated with the solution w_{B-1}. Show that it satisfies the time-update relations:

$$\xi(B) = \xi(B-1) + r_B^\mathsf{T}R_B^{-1}e_B \;=\; \xi(B-1) + e_B^\mathsf{T}T_Be_B, \quad \xi(-1) = 0$$

Conclude that $\xi(B) = \sum_{b=0}^{B} e_b^\mathsf{T}T_be_b$.

50.29 Consider the same formulation of Prob. 50.28 but assume the weighting matrix $\mathcal{R}_B$ is related to $\mathcal{R}_{B-1}$ as follows

$$\mathcal{R}_B = \begin{bmatrix} \mathcal{D}_{B-1}\mathcal{R}_{B-1} & \\ & R_B^{-1} \end{bmatrix}$$

where $\mathcal{D}_{B-1} = \text{diag}\{I_p, \ldots, I_p,\ \beta I_p,\ I_p, \ldots, I_p\}$, and $\beta > 1$ is a positive scalar. The scalar β appears at the location corresponding to the $k-$th block R_k^{-1}. Find a recursion relating w_B to w_{B-1}.

50.30 Consider a regularized block least-squares problem of the form

$$\min_{w\in\mathbb{R}^M} \left\{\lambda^{B+1}(w-\bar{w})^\mathsf{T}\Pi(w-\bar{w}) \;+\; \sum_{b=0}^{B}\lambda^{B-b}(d_b - U_bw)^\mathsf{T}R_b^{-1}(y_b - U_bw)\right\}$$

where each d_b has size $p \times 1$, each U_b has size $p \times M$, and each R_b is $p \times p$ and positive-definite. Moreover, $0 \ll \lambda \leq 1$ is an exponential forgetting factor and $\Pi > 0$. Let $\xi(B)$ denote the value of the minimum risk associated with the optimal solution w_B. Repeat the arguments of Prob. 50.28 to show that the solution w_B can be time-updated by the following block RLS algorithm:

$$\begin{cases} T_b &= (R_b + \lambda^{-1} U_b P_{b-1} U_b^{\mathsf{T}})^{-1} \\ G_b &= \lambda^{-1} P_{b-1} U_b^{\mathsf{T}} T_b \\ e_b &= d_b - U_b w_{b-1} \\ w_b &= w_{b-1} + G_b(d_b - U_b w_{b-1}), \quad w_{-1} = \bar{w} \\ P_b &= \lambda^{-1} P_{b-1} - G_b T_b^{-1} G_b^{\mathsf{T}}, \quad P_{-1} = \Pi^{-1} \\ r_b &= d_b - U_b w_b \\ \xi(b) &= \lambda \xi(b-1) + e_b^{\mathsf{T}} T_b e_b, \quad \xi(-1) = 0 \\ &= \lambda \xi(b-1) + r_b^{\mathsf{T}} R_b^{-1} e_b \end{cases}$$

Verify also that the quantities $\{G_b, T_b\}$ admit the alternative expressions $G_b = P_b U_b^{\mathsf{T}} R_b^{-1}$ and $T_b = R_b^{-1} - R_b^{-1} U_b P_b U_b^{\mathsf{T}} R_b^{-1}$.

50.31 Consider a collection of $N \times N$ symmetric matrices $\{A_m\}$ for $m = 1, 2, \ldots, M$, an $N \times N$ full-rank matrix U, and an $M \times 1$ vector b. It is assumed that $M \ll N^2$ so that the amount of data represented by the size of b is significantly smaller than the number of entries in U. Define the $M \times 1$ vector $\mathcal{A}(U) = \text{col}\{\text{Tr}(U^{\mathsf{T}} A_m U)\}$ and consider the optimization problem:

$$\min_{U \in \mathbb{R}^{N \times N}} \|\mathcal{A}(U) - b\|^2$$

Under $M \ll N^2$, there are many solutions U that satisfy $\mathcal{A}(U) = b$.

(a) Write down the gradient-descent recursion for seeking a minimizer for the above problem.

(b) Assume the matrices $\{A_m\}$ commute so that $A_m A_n = A_n A_m$ for any n and m. Argue that for a sufficiently small step-size, and for an initial condition close to zero, the gradient-descent algorithm converges towards the solution with the smallest nuclear norm, i.e., towards the solution U that solves

$$\min_{U} \|UU^{\mathsf{T}}\|_\star, \quad \text{subject to } \mathcal{A}(U) = b$$

Remark. The result of this problem provides another manifestation of the implicit bias/regularization problem discussed in the comments at the end of the Chapter. There are many solutions U for the over-parameterized problem; yet gradient descent converges to the solution with the smallest nuclear norm. See Gunasekar *et al.* (2017) for more discussion.

50.A MINIMUM-NORM SOLUTION

Let

$$\mathcal{W} = \{w \text{ such that } \|d - Hw\|^2 \text{ is minimum}\} \tag{50.173}$$

denote the set of all solutions to the standard least-squares problem (50.19). We argue below, motivated by the presentation from Sayed (2003,2008), that the solution to

$$\min_{w \in \mathcal{W}} \|w\| \tag{50.174}$$

is given by

$$w^\star = H^\dagger d \tag{50.175}$$

in terms of the pseudo-inverse of H.

Proof: We establish (50.175) for the over-determined case (i.e., when $N \geq M$) by introducing the singular-value decomposition (SVD) of H from Sec. 1.7. A similar argument applies to the under-determined case (when $N < M$). Thus, let $r \leq M$ denote the rank of H and introduce its SVD:

$$H = U \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] V^{\mathsf{T}} \tag{50.176}$$

where $\Sigma = \text{diag}\Big\{\sigma_1, \ldots, \sigma_r, 0, \ldots, 0\Big\}$. Then, it holds that

$$\|d - Hw\|^2 = \|U^{\mathsf{T}} d - U^{\mathsf{T}} H V V^{\mathsf{T}} w\|^2 \; = \; \left\| f - \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] z \right\|^2 \tag{50.177}$$

where we introduced the vectors $z = V^{\mathsf{T}} w$ and $f = U^{\mathsf{T}} d$. Note that z and w have the same Euclidean norm. Therefore, the problem of minimizing $\|d - Hw\|^2$ over w is equivalent to the problem of minimizing the rightmost term in (50.177) over z. Let $\{z(i), f(i)\}$ denote the individual entries of $\{z, f\}$. Then

$$\left\| f - \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] z \right\|^2 = \sum_{i=1}^{r} (f(i) - \sigma_i z(i))^2 \; + \; \sum_{i=r+1}^{N} f^2(i) \tag{50.178}$$

The second term is independent of z. Hence, any solution z has to satisfy $z(i) = f(i)/\sigma_i$ for $i = 1$ to r and $z(i)$ arbitrary for $i = r+1$ to $i = M$. The solution z with the smallest Euclidean norm requires that these latter values be set to zero. In this case, the solution becomes

$$\begin{aligned} w^{\star} &= V \; \text{col}\Big\{f(1)/\sigma_1, \ldots, f(r)/\sigma_r, 0, \ldots, 0\Big\} \\ &= V \left[\begin{array}{cc} \Sigma^{\dagger} & 0 \end{array} \right] U^{\mathsf{T}} d \; \overset{(1.115)}{=} \; H^{\dagger} d \end{aligned} \tag{50.179}$$

as claimed.

■

50.B EQUIVALENCE IN LINEAR ESTIMATION

There is a close relation between regularized least-squares problems and linear least-mean-squares estimation problems. Although the former class of problems deals with deterministic variables and the latter deals with random variables, both classes turn out to be equivalent in the sense that solving a problem from one class also solves a problem from the other class and vice-versa. We follow the presentation from Sayed and Kailath (1994), Kailath, Sayed, and Hassibi (2000), and Sayed (2003,2008).

Stochastic problem

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two zero-mean vector random variables that are related via a linear model of the form:

$$\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{v} \tag{50.180a}$$

for some known matrix H and where $\boldsymbol{v}$ denotes a zero-mean random noise vector with known covariance matrix, $R_v = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}$. The covariance matrix of $\boldsymbol{x}$ is also known and denoted by $\mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} = R_x$. Both $\{\boldsymbol{x}, \boldsymbol{v}\}$ are uncorrelated, i.e., $\mathbb{E}\,\boldsymbol{x}\boldsymbol{v}^{\mathsf{T}} = 0$, and we further assume that $R_x > 0$ and $R_v > 0$. We established in (29.95b) that the linear least-mean-squares estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\widehat{\boldsymbol{x}} = \left(R_x^{-1} + H^{\mathsf{T}} R_v^{-1} H\right)^{-1} H^{\mathsf{T}} R_v^{-1} \boldsymbol{y} \tag{50.180b}$$

and that the resulting minimum mean-square error matrix is

$$\text{m.m.s.e.} \;=\; \left(R_x^{-1} + H^{\mathsf{T}} R_v^{-1} H\right)^{-1} \tag{50.180c}$$

Deterministic problem

Now consider instead deterministic vector variables $\{x, y\}$ and a data matrix H relating them via

$$y = Hx + v \tag{50.181a}$$

where v denotes measurement noise. Assume further that we pose the problem of estimating x by solving the weighted regularized least-squares problem:

$$\min_{x} \left\{\; x^{\mathsf{T}} \Pi x \;+\; (y - Hx)^{\mathsf{T}} W (y - Hx) \;\right\} \tag{50.181b}$$

where $\Pi > 0$ is a regularization matrix and $W > 0$ is a weighting matrix. It is straightforward to verify by differentiation that the solution $\widehat{x}$ is given by

$$\widehat{x} = \left(\Pi + H^{\mathsf{T}} W H\right)^{-1} H^{\mathsf{T}} W y \tag{50.181c}$$

and that the resulting minimum cost is

$$\xi = y^{\mathsf{T}} \left(W^{-1} + H \Pi^{-1} H^{\mathsf{T}}\right)^{-1} y \tag{50.181d}$$

Equivalence

Expression (50.180b) provides the linear least-mean-squares estimator of $\boldsymbol{x}$ in a stochastic framework, while expression (50.181c) provides the least-squares estimate of x in a deterministic setting. It is clear that if we replace the quantities in (50.180b) by $R_x \longleftarrow \Pi^{-1}$ and $R_v \longleftarrow W^{-1}$, then the stochastic solution (50.180b) would coincide with the deterministic solution (50.181c). We therefore say that both problems are equivalent. Such equivalences play an important role in estimation and inference theories since they allow us to move back and forth between deterministic and stochastic formulations, and to determine the solution for one context from the solution to the other. Table 50.2 summarizes the relations between the variables in both domains. We consider one application of these equivalence results in the next appendix in the context of Kalman and smoothing filters.

50.C EXTENDED LEAST-SQUARES

If we refer to the derivation in Example 30.4 and examine the Kalman recursions in that context, we will find that they agree with the recursive least-squares recursions. In other words, the example shows that the growing memory ($\lambda = 1$) RLS algorithm is equivalent to a Kalman filter implementation for estimating an unknown model $\boldsymbol{x}_0 = \boldsymbol{w}$ from the observations.

Now model (30.102) is special and, therefore, the RLS filter is equivalent not to a full-blown Kalman filter but only to a special case of it — see Haykin *et al.* (1997) for another special case. In this appendix, following the equivalence approach of Sayed

Table 50.2 Equivalence of the stochastic and deterministic frameworks.

Stochastic setting	Deterministic setting
random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$	deterministic variables $\{x, y\}$
model $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{v}$	model $y = Hx + v$
covariance matrix, R_x	inverse regularization matrix, Π^{-1}
noise covariance, R_v	inverse weighting matrix W^{-1}
$\widehat{\boldsymbol{x}}$	$\widehat{x}$
$\min\limits_{K} \mathbb{E}\,(\boldsymbol{x} - K\boldsymbol{y})(\boldsymbol{x} - K\boldsymbol{y})^{\mathsf{T}}$	$\min\limits_{x} \left\{x^{\mathsf{T}}\Pi x + \|y - Hx\|^2_W\right\}$
$\widehat{\boldsymbol{x}} = \left(R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H\right)^{-1} H^{\mathsf{T}}R_v^{-1}\boldsymbol{y}$	$\widehat{x} = \left(\Pi + H^{\mathsf{T}}WH\right)^{-1} H^{\mathsf{T}}Wy$
m.m.s.e. $= \left(R_x^{-1} + H^{\mathsf{T}}R_v^{-1}H\right)^{-1}$	min. cost $= y^{\mathsf{T}}\left(W^{-1} + H\Pi^{-1}H^{\mathsf{T}}\right)^{-1} y$

and Kailath (1994) from the previous appendix, and adapting the presentation from Kailath, Sayed, and Hassibi (2000), we describe the general deterministic least-squares formulation that is equivalent to a full-blown Kalman filter. In so doing, we will arrive at the extended RLS algorithm (50.211), which is better suited for tracking the state of linear state-space models, as opposed to tracking the state of the special model (30.102), as is further illustrated in Sayed (2003,2008) by means of several special cases.

Deterministic estimation

Consider a collection of $(N+1)$ measurements $\{y_n\}$, possibly column vectors, that satisfy

$$y_n = H_n x_n + v_n \tag{50.182}$$

where the $\{x_n \in \mathbb{R}^M\}$ evolve in time according to the state recursion

$$x_{n+1} = F_n x_n + G_n u_n, \quad n \geq 0 \tag{50.183}$$

Here, the $\{F_n, G_n, H_n\}$ are known matrices and the $\{u_n, v_n\}$ denote disturbances or noises. Let further Π_0 be a positive-definite regularization matrix, and let $\{Q_n, R_n\}$ be positive-definite weighting matrices. Given the $\{y_n\}$, we pose the problem of estimating the *initial* state vector x_0 and the signals $\{u_0, u_1, \ldots, u_N\}$ in a regularized least-squares manner by solving

$$\min_{\{x_0, u_0, \ldots, u_N\}} \left\{ x_0^{\mathsf{T}}\Pi_0^{-1}x_0 \;+\; \sum_{n=0}^{N}(y_n - H_n x_n)^{\mathsf{T}}R_n^{-1}(y_n - H_n x_n) + \sum_{n=0}^{N} u_n^{\mathsf{T}}Q_n^{-1}u_n \right\} \tag{50.184}$$

subject to the constraint (50.183). We denote the solution by $\{\widehat{x}_{0|N}, \widehat{u}_{n|N}, 0 \leq n \leq N\}$, and we refer to them as *smoothed* estimates since they are based on observations beyond the times of occurrence of the respective variables $\{x_0, u_n\}$.

In principle, we could solve (50.184) by using optimization arguments, e.g., based on the use of Lagrange multipliers. Instead, we will solve it by appealing to the equivalence result of Table 50.2. In other words, we will first determine the equivalent stochastic problem and then solve this latter problem to arrive at the solution of (50.184). This method of solving (50.184) not only serves as an illustration of the convenience of equivalence results in estimation theory, but it also shows that sometimes it is easier to solve a deterministic problem in the stochastic domain (or vice-versa). In our case, the problem at hand is more conveniently solved in the stochastic domain.

Introduce the column vectors

$$z \;\stackrel{\Delta}{=}\; \begin{bmatrix} x_0 \\ \hline u_0 \\ u_1 \\ \vdots \\ u_N \end{bmatrix}, \quad d \;\stackrel{\Delta}{=}\; \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{50.185}$$

as well as the block-diagonal matrices

$$\mathcal{W}^{-1} \;\stackrel{\Delta}{=}\; \text{blkdiag}\Big\{R_0, R_1, \ldots, R_N\Big\}, \qquad \Pi^{-1} \;\stackrel{\Delta}{=}\; \text{blkdiag}\Big\{\Pi_0, Q_0, \ldots, Q_N\Big\} \tag{50.186}$$

Then, it holds that

$$x_0^{\mathsf{T}}\Pi_0^{-1}x_0 \;+\; \sum_{n=0}^{N} u_n^{\mathsf{T}}Q_n^{-1}u_n = z^{\mathsf{T}}\Pi z \tag{50.187}$$

Moreover, by using the state equation (50.183) to express each term $H_n x_n$ in terms of combinations of the entries of z, we can verify that

$$\sum_{n=0}^{N}(y_n - H_n x_n)^{\mathsf{T}}R_n^{-1}(y_n - H_n x_n) = (d - \mathcal{H}z)^{\mathsf{T}}\mathcal{W}(d - \mathcal{H}z) \;=\; \|d - \mathcal{H}z\|^2_{\mathcal{W}} \tag{50.188}$$

where the matrix $\mathcal{H}$ is block lower-triangular and given by

$$\mathcal{H} \;\stackrel{\Delta}{=}\; \begin{bmatrix} H_0 & & & & & \\ H_1\Phi(1,0) & H_1G_0 & & & & \\ H_2\Phi(2,0) & H_2\Phi(2,1)G_0 & H_2G_1 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ H_N\Phi(N,0) & H_N\Phi(N,1)G_0 & H_N\Phi(N,2)G_1 & \ldots & H_NG_{N-1} & 0 \end{bmatrix} \tag{50.189}$$

and the matrices $\Phi(n,m)$ are defined by

$$\Phi(n,m) \;\stackrel{\Delta}{=}\; \begin{cases} F_{n-1}F_{n-2}\ldots F_m, & n > m \\ I_M, & n = m \end{cases} \tag{50.190}$$

In other words, we find that we can rewrite the original cost function (50.184) as the regularized least-squares problem:

$$\min_{z} \Big\{z^{\mathsf{T}}\Pi z \;+\; (d - \mathcal{H}z)^{\mathsf{T}}\mathcal{W}(d - \mathcal{H}z)\Big\} \tag{50.191}$$

Let $\widehat{z}_N$ denote the solution to (50.191), i.e., $\widehat{z}_N$ is a column vector that contains the desired solutions:

$$\widehat{z}_N = \text{col}\Big\{\widehat{x}_{0|N}, \widehat{u}_{0|N}, \widehat{u}_{1|N}, \ldots, \widehat{u}_{N|N}\Big\} \tag{50.192}$$

Now, in view of the equivalence result from Table 50.2, we know that $\widehat{z}_N$ can be obtained by solving an equivalent stochastic estimation problem that is determined as follows.

Stochastic estimation

We introduce zero-mean random vectors $\{\boldsymbol{z}, \boldsymbol{d}\}$, with the same dimensions and partitioning as the above $\{z, d\}$, and assume that they are related via a linear model of the form:

$$\boldsymbol{d} = \mathcal{H}\boldsymbol{z} + \boldsymbol{v} \tag{50.193}$$

where $\mathcal{H}$ is the same matrix as in (50.189), and where $\boldsymbol{v}$ denotes a zero-mean additive noise vector, uncorrelated with $\boldsymbol{z}$, and partitioned as $\boldsymbol{v} = \text{col}\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$. The dimensions of the $\{\boldsymbol{v}_n\}$ are compatible with those of $\{\boldsymbol{y}_n\}$. We denote the covariance matrices of $\{\boldsymbol{z}, \boldsymbol{v}\}$ by

$$R_z = \mathbb{E}\,\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}, \qquad R_v = \mathbb{E}\,\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} \tag{50.194}$$

and we choose them as $R_z = \Pi^{-1}$ and $R_v = \mathcal{W}^{-1}$, where $\{\Pi, \mathcal{W}\}$ are given by (50.186).

Let $\widehat{\boldsymbol{z}}_N$ denote the linear least-mean-square error (l.l.m.s.e.) estimator of $\boldsymbol{z}$ given $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ in $\boldsymbol{d}$. We partition $\boldsymbol{z}$ as

$$\boldsymbol{z} = \text{col}\{\boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\} \tag{50.195}$$

Then the equivalence result of Table 50.2 states that the expression for $\widehat{\boldsymbol{z}}_{|N}$ in terms of $\boldsymbol{d}$ in the stochastic setting (50.193) is identical to the expression for $\widehat{z}_{|N}$ in terms of d in the deterministic problem (50.191).

In order to determine $\widehat{\boldsymbol{z}}_N$ or, equivalently, $\{\widehat{\boldsymbol{x}}_{0|N}, \widehat{\boldsymbol{u}}_{n|N}\}$, we start by noting that the linear model (50.193), coupled with the definitions of $\{R_z, R_v, \mathcal{H}\}$ in (50.186), (50.189), and (50.194), show that the stochastic variables $\{\boldsymbol{y}_n, \boldsymbol{v}_n, \boldsymbol{x}_0, \boldsymbol{u}_n\}$ so defined satisfy the following state-space model:

$$\begin{aligned} \boldsymbol{x}_{n+1} &= F_n\boldsymbol{x}_n + G_n\boldsymbol{u}_n \\ \boldsymbol{y}_n &= H_n\boldsymbol{x}_n + \boldsymbol{v}_n \end{aligned} \tag{50.196}$$

with

$$\mathbb{E}\begin{bmatrix} \boldsymbol{u}_n \\ \boldsymbol{v}_n \\ \boldsymbol{x}_0 \\ 1 \end{bmatrix}\begin{bmatrix} \boldsymbol{u}_m \\ \boldsymbol{v}_m \\ \boldsymbol{x}_0 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} Q_n\delta_{nm} & 0 & 0 \\ 0 & R_n\delta_{nm} & 0 \\ 0 & 0 & \Pi_0 \\ 0 & 0 & 0 \end{bmatrix} \tag{50.197}$$

We now use this model to derive recursions for estimating $\boldsymbol{z}$ (i.e., for estimating the variables $\{\boldsymbol{x}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$).

Solving the stochastic problem

Let $\widehat{\boldsymbol{z}}_n$ denote the l.l.m.s.e. estimator of $\boldsymbol{z}$ given the top entries $\{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_n\}$ in $\boldsymbol{d}$. To determine $\widehat{\boldsymbol{z}}_n$, and ultimately $\widehat{\boldsymbol{z}}_N$, we proceed recursively by employing the innovations $\{\boldsymbol{e}_n\}$ of the observations $\{\boldsymbol{y}_n\}$. Using the basic recursive estimation formula (30.23) we have

$$\begin{aligned} \widehat{\boldsymbol{z}}_n &= \widehat{\boldsymbol{z}}_{n-1} + (\mathbb{E}\,\boldsymbol{z}\boldsymbol{e}_n^{\mathsf{T}})\,R_{e,n}^{-1}\,\boldsymbol{e}_n \\ &= \widehat{\boldsymbol{z}}_{n-1} + \left(\mathbb{E}\,\boldsymbol{z}\widetilde{\boldsymbol{x}}_{n|n-1}^{\mathsf{T}}\right)H_n^{\mathsf{T}}R_{e,n}^{-1}\,\boldsymbol{e}_n, \quad \widehat{\boldsymbol{z}}_{-1} = 0 \end{aligned} \tag{50.198}$$

where we used in the second equality the innovations equation (cf. (30.51)):

$$\boldsymbol{e}_n = \boldsymbol{y}_n - H_n\widehat{\boldsymbol{x}}_{n|n-1} \;=\; H_n\widetilde{\boldsymbol{x}}_{n|n-1} + \boldsymbol{v}_n \tag{50.199}$$

and the fact that $\mathbb{E}\,\boldsymbol{x}_0\boldsymbol{v}_m^{\mathsf{T}} = 0$ and $\mathbb{E}\,\boldsymbol{u}_n\boldsymbol{v}_m^{\mathsf{T}} = 0$ for all m. Clearly, the entries of $\widehat{\boldsymbol{z}}_n$ have the interpretation

$$\widehat{\boldsymbol{z}}_n = \text{col}\Big\{\widehat{\boldsymbol{x}}_{0|n}, \widehat{\boldsymbol{u}}_{0|n}, \widehat{\boldsymbol{u}}_{1|n}, \ldots, \widehat{\boldsymbol{u}}_{n-1|n}, 0, 0, \ldots, 0\Big\} \tag{50.200}$$

where the trailing entries of $\widehat{\boldsymbol{z}}_n$ are zero since $\widehat{\boldsymbol{u}}_{m|n} = 0$ for $m \geq n$.

Let $K_{z,n} = \mathbb{E}\,\boldsymbol{z}\widetilde{\boldsymbol{x}}^{\mathsf{T}}_{n|n-1}$. The above recursive construction would be complete, and hence provide the desired quantity $\widehat{\boldsymbol{z}}_N$, once we show how to evaluate the gain matrix $K_{z,n}$. For this purpose, we first subtract the equations (from the Kalman filter (30.69)):

$$\boldsymbol{x}_{n+1} = F_n\boldsymbol{x}_n + G_n\boldsymbol{u}_n \tag{50.201}$$

$$\widehat{\boldsymbol{x}}_{n+1|n} = F_n\widehat{\boldsymbol{x}}_{n|n-1} + K_{p,n}(H_n\widetilde{\boldsymbol{x}}_{n|n-1} + \boldsymbol{v}_n) \tag{50.202}$$

to obtain

$$\widetilde{\boldsymbol{x}}_{n+1|n} = F_{p,n}\widetilde{\boldsymbol{x}}_{n|n-1} + G_n\boldsymbol{u}_n - K_{p,n}\boldsymbol{v}_n \tag{50.203}$$

where $F_{p,n} = F_n - K_{p,n}H_n$. Using this recursion, it is easy to verify that $K_{z,n}$ satisfies the recursion:

$$K_{z,n+1} \overset{\Delta}{=} \mathbb{E}\,\boldsymbol{z}\widetilde{\boldsymbol{x}}^{\mathsf{T}}_{n+1|n} = K_{z,n}F^{\mathsf{T}}_{p,n} + \left[\begin{array}{c} 0 \\ \hline 0 \\ I \\ 0 \end{array}\right] Q_n G_n^{\mathsf{T}}, \quad K_{z,0} = \left[\begin{array}{c} \Pi_0 \\ 0 \end{array}\right] \tag{50.204}$$

The identity matrix that appears in the second term of the recursion for $K_{z,n+1}$ occurs at the position that corresponds to the entry $\boldsymbol{u}_n$ in the vector $\boldsymbol{z}$, e.g.,

$$K_{z,1} = \begin{bmatrix} \Pi_0 F^{\mathsf{T}}_{p,0} \\ Q_0 G_0^{\mathsf{T}} \\ 0 \end{bmatrix}, \qquad K_{z,2} = \begin{bmatrix} \Pi_0 F^{\mathsf{T}}_{p,0}F^{\mathsf{T}}_{p,1} \\ Q_0 G_0^{\mathsf{T}} F^{\mathsf{T}}_{p,1} \\ Q_1 G_1^{\mathsf{T}} \\ 0 \end{bmatrix}, \quad \ldots \tag{50.205}$$

Substituting (50.204) into (50.198) we find that the following recursions hold:

$$\left\{\begin{array}{lcl} \widehat{\boldsymbol{x}}_{0|n} & = & \widehat{\boldsymbol{x}}_{0|n-1} + \Pi_0\Phi_p^{\mathsf{T}}(n,0)H_n^{\mathsf{T}}R_{e,n}^{-1}\boldsymbol{e}_n, \quad \widehat{\boldsymbol{x}}_{0|-1} = 0 \\ \widehat{\boldsymbol{u}}_{m|n} & = & \widehat{\boldsymbol{u}}_{m|n-1} + Q_m G_m^{\mathsf{T}}\Phi_p^{\mathsf{T}}(n,m+1)H_n^{\mathsf{T}}R_{e,n}^{-1}\boldsymbol{e}_n, \quad m < n \\ \widehat{\boldsymbol{u}}_{m|n} & = & 0, \;\; m \geq n \end{array}\right. \tag{50.206}$$

where the matrix $\Phi_p(n,m)$ is defined by

$$\Phi_p(n,m) \overset{\Delta}{=} \left\{\begin{array}{ll} F_{p,n-1}F_{p,n-2}\ldots F_{p,m}, & n > m \\ I, & m = n \end{array}\right. \tag{50.207}$$

If we introduce the auxiliary variable

$$\boldsymbol{\lambda}_{n|N} \overset{\Delta}{=} \sum_{m=n}^{N} \Phi_p^{\mathsf{T}}(m,n)H_m^{\mathsf{T}}R_{e,m}^{-1}\boldsymbol{e}_m \tag{50.208}$$

then it is easy to verify that recursions (50.206) lead to

$$\left\{\begin{array}{lcl} \widehat{\boldsymbol{x}}_{0|N} & = & \Pi_0\boldsymbol{\lambda}_{0|N} \\ \widehat{\boldsymbol{x}}_{m+1|m} & = & F_{p,m}\widehat{\boldsymbol{x}}_{m|m-1} + K_{p,m}\boldsymbol{y}_m, \quad \widehat{\boldsymbol{x}}_{0|-1} = 0 \\ \boldsymbol{e}_m & = & \boldsymbol{y}_m - H_m\widehat{\boldsymbol{x}}_{m|m-1} \\ \widehat{\boldsymbol{u}}_{m|N} & = & Q_m G_m^{\mathsf{T}}\boldsymbol{\lambda}_{m+1|N} \\ \boldsymbol{\lambda}_{m|N} & = & F^{\mathsf{T}}_{p,m}\boldsymbol{\lambda}_{m+1|N} + H_m^{\mathsf{T}}R_{e,m}^{-1}\boldsymbol{e}_m, \quad \boldsymbol{\lambda}_{N+1|N} = 0 \end{array}\right. \tag{50.209}$$

These equations are the Bryson-Frazier smoothing recursions (30.194) — refer also to Prob. 30.13; the recursions (30.194) evaluate the estimators $\{\widehat{\boldsymbol{x}}_{0|n}, \widehat{\boldsymbol{u}}_{m|n}\}$ for successive values of n, and not only for $n = N$ as in (50.209). Just like $\{\widehat{\boldsymbol{x}}_{0|N}, \widehat{\boldsymbol{u}}_{n|N}\}$, the estimators $\{\widehat{\boldsymbol{x}}_{0|n}, \widehat{\boldsymbol{u}}_{m|n}\}$ can also be related to the solution of a least-squares problem. Indeed, by equivalence, the expressions that provide the solutions $\{\widehat{\boldsymbol{x}}_{0|n}, \widehat{\boldsymbol{u}}_{m|n}\}$ in (50.206) should coincide with those that provide the solutions $\{\widehat{x}_{0|n}, \widehat{u}_{m|n}\}$ for the following deterministic problem, with data up to time n (rather than N as in (50.184)):

$$\min_{x_0,u_0,\ldots,u_n} \left\{ x_0^{\mathsf{T}}\Pi_0^{-1}x_0 \;+\; \sum_{m=0}^{n}(y_m - H_m x_m)^{\mathsf{T}} R_m^{-1}(y_m - H_m x_m) + \sum_{m=0}^{n} u_m^{\mathsf{T}} Q_m^{-1} u_m \right\} \tag{50.210}$$

We know by equivalence that the mapping from $\{\boldsymbol{y}_m\}$ to $\{\widehat{\boldsymbol{x}}_{0|N}, \widehat{\boldsymbol{u}}_{m|N}\}$ in the stochastic problem (50.193) coincides with the mapping from $\{y_m\}$ to $\{\widehat{x}_{0|N}, \widehat{u}_{m|N}\}$ in the deterministic problem (50.191). We are therefore led to listing (50.211).

Extended recursive least-squares algorithm to solve (50.184)

```
given observations $\{y_n\}$ that satisfy $x_{n+1} = F_n x_n + G_n u_n$
      and $y_n = H_n x_n + v_n$;
objective: estimate $\{x_0, u_0, u_1, \ldots, u_n\}$ by solving (50.184).
start from $\widehat{x}_{0|-1} = 0, P_{0|-1} = \Pi_0, \lambda_{N+1|N} = 0$.
(forward pass)
repeat for $n = 0, 1, 2, \ldots$:
    $e_n = y_n - H_n \widehat{x}_{n|n-1}$
    $R_{e,n} = R_n + H_n P_{n|n-1} H_n^{\mathsf{T}}$
    $K_{p,n} = F_n P_{n|n-1} H_n^{\mathsf{T}} R_{e,n}^{-1}$
    $\widehat{x}_{n+1|n} = F_n \widehat{x}_{n|n-1} + K_{p,n} e_n$                                  (50.211)
    $P_{n+1|n} = F_n P_{n|n-1} F_n^{\mathsf{T}} + G_n Q_n G_n^{\mathsf{T}} - K_{p,n} R_{e,n} K_{p,n}^{\mathsf{T}}$
end
(backward pass)
repeat for $n = N, N-1, \ldots, 1, 0$:
    $F_{p,n} = F_n - K_{p,n} H_n$
    $\lambda_{n|N} = F_{p,n}^{\mathsf{T}} \lambda_{n+1|N} \;+\; H_n^{\mathsf{T}} R_{e,n}^{-1} e_n$
end
(output)
set $\widehat{x}_{0|N} = \Pi_0 \lambda_{0|N}$
set $\widehat{u}_{n|N} = Q_n G_n^{\mathsf{T}} \lambda_{n+1|N}, \; 0 \le n \le N$
```

REFERENCES

Adrain, R. (1808), "Research concerning the probabilities of the errors which happen in making observations," *The Analyst*, vol. I, no. 4, pp. 93–109.

Ailon, N. and E. Liberty (2013), "Almost optimal unrestricted fast Johnson–Lindenstrauss transform," *ACM Transactions on Algorithms*, vol. 9, no. 3, article 21, available online at https://doi.org/10.1145/2483699.2483701.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, 3rd edition, Wiley, NY.

Anderson, B. D. O. and J. B. Moore (1979), *Optimal Filtering*, Prentice Hall, NJ.

Aström, K. J. and B. Wittenmark (1971), "Problems of identification and control," *J. Math. Anal. App.*, vol. 34, pp. 90–113.

Beckley, B. D., P. S. Callahan, D. W. Hancock, G. T. Mitchum, and R. D. Ray (2017), "On the cal-mode correction to TOPEX satellite altimetry and its effect on the global mean sea level time series," *J. Geophysical Research: Oceans*, vol. 122, no. 11, pp. 8371–8384.

Belkin, M., S. Ma, and S. Mandal (2018), "To understand deep learning we need to understand kernel learning," *available online at arXiv:1802.01396.*

Belkin, M., A. Rakhlin, and A. B. Tsybakov (2019), "Does data interpolation contradict statistical optimality?" *Proc. International Conference on Artificial Intelligence and Statistics* (AISTATS), pp. 1611–1619, Naha, Japan.

Bell, R. and Y. Koren (2007a), "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," *Proc. IEEE International Conference on Data Mining* (ICDM), pp. 43–52, Omaha, NE.

Bjorck, A. (1996), *Numerical Methods for Least Squares Problems,* SIAM, PA.

Businger, P. and G. H. Golub (1965), "Linear least-squares solution by Householder transformations," *Numer. Math.*, vol. 7, pp. 269–276.

Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing scatterplots," *J. American Statistical Association*, vol. 74, pp. 829–836.

Cleveland, W. S. and Devlin, S. J. (1988), "Locally weighted regression: An approach to regression analysis by local fitting," *J. American Statistical Association*, vol. 83, pp. 596–610.

Draper, N. R. and H. Smith (1998), *Applied Regression Analysis*, 3nd edition, Wiley, NY.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2006b), "Subspace sampling and relative–error matrix approximation: Column–row-based methods," in *Proc. Algorithms — Annual European Symposium* (ESA), pp. 304–314, Zurich, Switzerland.

Eaton, M. L. (1983), *Multivariate Statistics: A Vector Space Approach*, Wiley, NY.

Fletcher, R., J. A. Grant, and M. D. Hebden (1971), "The calculation of linear best L_p approximations," *Comput. J.*, vol. 14, pp. 276–279.

Gauss, C. F. (1809), *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, English translation by C. H. Davis, 1857, Little, Brown, and Company, Boston, MA.

Godard, D. N. (1974), "Channel equalization using a Kalman filter for fast data transmission," *IBM J. Res. Develop.*, vol. 18, pp. 267–273.

Golub, G. H. (1965), "Numerical methods for solving linear least-squares problems," *Numer. Math.*, vol. 7, pp. 206–216.

Gower, R. M. and P. Richtárik (2015), "Randomized iterative methods for linear systems," *SIAM J. Matrix Analysis and Applications*, vol. 36, no. 4, pp. 1660–1690.

GSFC (2017), "Global mean sea level trend from integrated multi-mission ocean altimeters TOPEX/Poseidon, Jason-1, OSTM/Jason-2," ver. 4.2 PO.DAAC, CA, USA. Dataset accessed 2019-03-18 at http://dx.doi.org/10.5067/GMSLM-TJ42.

Gunasekar, S., J. Lee, D. Soudry, and N. Srebro (2018), "Characterizing implicit bias in terms of optimization geometry," *in Proc. International Conference on Machine Learning* (ICML), pp. 1832–1841, Stockholm, Sweden.

Gunasekar, S., B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro (2017), "Implicit regularization in matrix factorization," *Proc. Neural Information Processing Systems* (NIPS), pp. 6151–6159, Long Beach, CA.

Gupta, A. K. and D. K. Nagar (2000), *Matrix Variate Distributions*, Chapman & Hall.

Hall, T. (1970), *Carl Friedrich Gauss: A Biography*, MIT Press, Cambridge, MA.

Hastie, T., A. Montanari, S. Rosset, and R. Tibshirani (2019), "Surprises in high-dimensional ridgeless least squares interpolation," *available online at arXiv:1903.08560*

Haykin, S. (1991), *Adaptive Filter Theory*, 2nd edition, Prentice Hall, NJ.

Haykin, S., A. H. Sayed, J. Zeidler, P. Wei, and P. Yee (1997), "Adaptive tracking of linear time-variant systems by extended RLS algorithms," *IEEE Trans. Signal Processing*, vol. 45, no. 5, pp. 1118–1128.

Higham, N. J. (1996), *Accuracy and Stability of Numerical Algorithms*, SIAM, PA.

Ho, Y. C. (1963), "On the stochastic approximation method and optimal filter theory," *J. Math. Anal. Appl.*, vol. 6, pp. 152–154.

Householder, A. S. (1953), *Principles of Numerical Analysis,* McGraw-Hill, NY.

Hu, Y. F., Y. Koren, and C. Volinsky (2008), "Collaborative filtering for implicit feedback datasets," *Proc. IEEE International Conference on Data Mining* (ICDM), pp. 263–272, Pisa, Italy.

Indyk, P. and R. Motwani (1998), "Approximate nearest neighbors: Towards removing the curse of dimensionality," *Proc. ACM Symposium on Theory of Computing*, pp. 604–613, Dallas, TX.

Jin, H. and G. Montúfar (2020), " Implicit bias of gradient descent for mean squared error regression with wide neural networks," *available online at arXiv:2006.07356.*

Johnson, W. and J. Lindenstrauss (1984), "Extensions of Lipschitz maps into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206.

Kahng, S. W. (1972), "Best L_p approximation," *Math. Comput.*, vol. 26, pp. 505–508.

Kailath, T., A. H. Sayed, and B. Hassibi (2000), *Linear Estimation*, Prentice Hall, NJ.

Lawson, C. L. and R. J. Hanson (1995), *Solving Least-Squares Problems*, SIAM, PA.

Legendre, A. M. (1805), *Nouvelles Méthodes pour la Détermination des Orbites de Comètes*, Courcier, Paris.

Legendre, A. M. (1810), "Méthode de moindres quarres, pour trouver le milieu de plus probable entre les résultats des différentes observations," *Mem. Inst. France*, pp. 149–154.

Ljung, L. (1987), *System Identification: Theory for the User*, Prentice Hall, NJ.

Mahoney, M. W. (2011), *Randomized Algorithms for Matrices and Data*, Foundations and Trends in Machine Learning, NOW Publishers, vol. 3, no. 2, pp. 123–224.

McClave, J. T. and T. T. Sincich (2016), *Statistics*, 13th edition, Pearson.

Mei, S. and A Montanari (2020), "The generalization error of random features regression: Precise asymptotics and double descent curve," *available at arXiv:1908.05355.*

Mendenhall, W., R. J. Beaver, and B. M. Beaver (2012), *Introduction to Probability and Statistics*, 14th edition, Cenage Learning.

Neyshabur, B., R. Tomioka, and N. Srebro (2015), "In search of the real inductive bias: On the role of implicit regularization in deep learning," *available online at https://arxiv.org/abs/1412.6614.*

Osborne, M. R. (1985), *Finite Algorithms in Optimization and Data Analysis*, Wiley, NY.

Pilanci, M. and M. J. Wainwright (2015), "Randomized sketches of convex programs With sharp guarantees," *IEEE Trans. Information Theory*, vol. 61, no. 9, pp. 5096–5115.

Pilaszy, I., D. Zibriczky, and D. Tikk (2010), "Fast ALS-based matrix factorization for explicit and implicit feedback datasets," *Proc. ACM conference on Recommender Systems*, pp. 71–78.

Plackett, R. L. (1950), "Some theorems in least-squares," *Biometrika*, vol. 37, no. 1–2, pp. 149–157.

Plackett, R. L. (1972), "The discovery of the method of least-squares," *Biometrika*, vol. 59, pp. 239–251.

Rigollet, P. and J.-C. Huetter (2017), *High Dimensional Statistics*, MIT Lecture Notes, available online at http://www-math.mit.edu/ rigollet/PDFs/RigNotes17.pdf

Sarlós, T. (2006), "Improved approximation algorithms for large matrices via random projections," *Proc. IEEE Symposium on Foundations of Computer Science* (FOCS), pp. 143–152, Berkeley, CA.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley, NJ.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley, NJ.

Sayed, A. H. and T. Kailath (1994), "A state-space approach to adaptive RLS filtering," *IEEE Signal Processing Magazine*, vol. 11, no. 3, pp. 18–60.

Söderström, T. (1994), *Discrete-Time Stochastic Systems: Estimation and Control*, Prentice-Hall International, UK.

Sorenson, H. W. (1966), "Kalman filtering techniques," in *Advances in Control Systems Theory and Applications*, C. T. Leondes, *Editor*, vol. 3, pp. 219–292, Academic Press.

Soudry, D., E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro (2018), "The implicit bias of gradient descent on separable data," *J. Machine Learning Research*, vol. 19, pp. 1–57.

Stewart, G. W. (1995), *Theory of the Combination of Observations Least Subject to Errors*, Classics in Applied Mathematics, SIAM, PA. [Translation of original works by C. F. Gauss under the title *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.*]

Stigler, S. M. (1981), "Gauss and the invention of least-squares," *The Annals of Statistics*, vol. 9, no. 3, pp. 465–474.

Strobach, P. (1990), *Linear Prediction Theory*, Springer-Verlag, Berlin Heidelberg.

Udell, M., C. Horn, R. Zadeh, and S. Boyd (2016), *Generalized Low Rank Models*, vol. 9, no. 1, pp. 1–118, NOW Publishers.

Willsky, A. S. (1979), *Digital Signal Processing and Control and Estimation Theory*, MIT Press, Cambridge, MA.

Witte, R. S. and J. S. Witte (2013), *Statistics*, 10th edition, Wiley, NY.

Woodruff, D. P. (2014), *Sketching as a Tool for Numerical Linear Algebra*, Foundations and Trends in Theoretical Computer Science, NOW Publishers, vol. 10, no. 1–2, pp. 1–157.

Zhou, Y., D. Wilkinson, R. Schreiber, and R. Pan (2008), "Large-scale parallel collaborative filtering for the Netflix prize," in *Algorithmic Aspects in Information and Management*, pp. 337–348, Springer, NY.