

Recursive Least Squares Estimation*

(Com 477/577 Notes)

Yan-Bin Jia

Dec 3, 2019

1 Estimation of a Constant

We start with estimation of a constant based on several noisy measurements. Suppose we have a resistor but do not know its resistance. So we measure it several times using a cheap (and noisy) multimeter. How do we come up with a good estimate of the resistance based on these noisy measurements?

More formally, suppose $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ is a constant but unknown vector, and $\mathbf{y} = (y_1, y_2, \dots, y_l)^T$ is an l -element noisy measurement vector. Our task is to find the “best” estimate $\tilde{\mathbf{x}}$ of \mathbf{x} . Here we look at perhaps the simplest case where each y_i is a linear combination of x_j , $1 \leq j \leq n$, with addition of some measurement noise ν_i . Thus, we are working with the following linear system,

$$\mathbf{y} = H\mathbf{x} + \boldsymbol{\nu},$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \dots, \nu_l)^T$, and H is an $l \times n$ matrix; or with all terms listed,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} = \begin{pmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & \ddots & \vdots \\ H_{l1} & \cdots & H_{ln} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_l \end{pmatrix}.$$

Given an estimate $\tilde{\mathbf{x}}$, we consider the difference between the noisy measurements and the projected values $H\tilde{\mathbf{x}}$:

$$\boldsymbol{\epsilon} = \mathbf{y} - H\tilde{\mathbf{x}}.$$

Under the least squares principle, we will try to find the value of $\tilde{\mathbf{x}}$ that minimizes the cost function

$$\begin{aligned} J(\tilde{\mathbf{x}}) &= \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &= (\mathbf{y} - H\tilde{\mathbf{x}})^T (\mathbf{y} - H\tilde{\mathbf{x}}) \\ &= \mathbf{y}^T \mathbf{y} - \tilde{\mathbf{x}}^T H \mathbf{y} - \mathbf{y}^T H \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T H^T H \tilde{\mathbf{x}}. \end{aligned}$$

The necessary condition for the minimum is the vanishing of the partial derivative of J with respect to $\tilde{\mathbf{x}}$, that is,

$$\frac{\partial J}{\partial \tilde{\mathbf{x}}} = -2\mathbf{y}^T H + 2\tilde{\mathbf{x}}^T H^T H = 0.$$

*The material is adapted from Sections 3.1–3.3 in Dan Simon’s book *Optimal State Estimation* [1].

We solve the equation, obtaining

$$\tilde{\mathbf{x}} = (H^T H)^{-1} H^T \mathbf{y}. \quad (1)$$

The inverse $(H^T H)^{-1}$ exists if $\text{rank}(H) = n$ (which implies $l \geq n$). In other words, when the number of measurements is no fewer than the number of variables, and these measurements are linearly independent.

EXAMPLE 1. Suppose we are trying to estimate the resistance x of an unmarked resistor based on l noisy measurements using a multimeter. In this case,

$$\mathbf{y} = Hx + \boldsymbol{\nu}, \quad (2)$$

where

$$H = (1, \dots, 1)^T. \quad (3)$$

Substitution of the above into equation (1) gives us the optimal estimate of x as

$$\begin{aligned} \tilde{x} &= (H^T H)^{-1} H^T \mathbf{y} \\ &= \frac{1}{l} H^T \mathbf{y} \\ &= \frac{y_1 + \dots + y_l}{l}. \end{aligned} \quad (4)$$

2 Weighed Least Squares Estimation

So far we have placed equal confidence on all the measurements. Now we look at varying confidence in the measurements. For instance, some of our measurements of an unmarked resistor were taken with an expensive multimeter with low noise, while others were taken with a cheap multimeter by a tired student late at night. Even though the second set of measurements is less reliable, we could get some information about the resistance. We should never throw away measurements, no matter how unreliable they may seem.

We assume that each measurement y_i , $1 \leq i \leq l$, may be taken under a different condition so that the variance ν_i of the measurement noise may be distinct too:

$$E(\nu_i^2) = \sigma_i^2, \quad 1 \leq i \leq l.$$

We also assume that the noise for each measurement has zero mean and is independent. The covariance matrix for all measurement noise is

$$\begin{aligned} R &= E(\boldsymbol{\nu}\boldsymbol{\nu}^T) \\ &= \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l^2 \end{pmatrix}. \end{aligned}$$

Writing the difference $\mathbf{y} - H\tilde{\mathbf{x}}$ as $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_l)^T$, we will minimize the sum of squared differences weighted over the variations of the measurements:

$$J(\tilde{\mathbf{x}}) = \boldsymbol{\epsilon}^T R^{-1} \boldsymbol{\epsilon} = \frac{\epsilon_1^2}{\sigma_1^2} + \frac{\epsilon_2^2}{\sigma_2^2} + \dots + \frac{\epsilon_l^2}{\sigma_l^2}.$$

If a measurement y_i is noisy (as indicated by a large standard deviation σ_i), we care less about the discrepancy between it and the i th element of $H\tilde{\mathbf{x}}$ because we do not have much confidence in this measurement. The cost function J can be expanded as follows:

$$\begin{aligned} J(\tilde{\mathbf{x}}) &= (\mathbf{y} - H\tilde{\mathbf{x}})^T R^{-1} (\mathbf{y} - H\tilde{\mathbf{x}}) \\ &= \mathbf{y}^T R^{-1} \mathbf{y} - \tilde{\mathbf{x}}^T H^T R^{-1} \mathbf{y} - \mathbf{y}^T R^{-1} H \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T H^T R^{-1} H \tilde{\mathbf{x}}. \end{aligned}$$

At a minimum, the partial derivative of J must vanish, yielding

$$\frac{\partial J}{\partial \tilde{\mathbf{x}}} = -2\mathbf{y}^T R^{-1} H + 2\tilde{\mathbf{x}}^T H^T R^{-1} H = 0.$$

Immediately, we solve the above equation for the best estimate of \mathbf{x} :

$$\tilde{\mathbf{x}} = (H^T R^{-1} H)^{-1} H^T R^{-1} \mathbf{y}. \quad (5)$$

Note that the measurement noise matrix R must be non-singular for a solution to exist. In other words, each measurement y_i must be corrupted by some noise for the estimation method to work.

EXAMPLE 2. We get back to the problem in Example 1 of resistance estimation, for which the equations are given in (2) and (3). Suppose each of the l noisy measurements has variance

$$E(\nu_i^2) = \sigma_i^2.$$

The measurement noise covariance is given as

$$R = \text{diag}(\sigma_1^2, \dots, \sigma_l^2).$$

Substituting H, R, \mathbf{y} into (5), we obtain the estimate

$$\begin{aligned} \tilde{\mathbf{x}} &= \begin{pmatrix} (1, \dots, 1) \end{pmatrix} \begin{pmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_l^2 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}^{-1} \begin{pmatrix} (1, \dots, 1) \end{pmatrix} \begin{pmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_l^2 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} \\ &= \left(\sum_{i=1}^l \frac{1}{\sigma_i^2} \right)^{-1} \begin{pmatrix} \sum_{i=1}^l \frac{y_i}{\sigma_i^2} \end{pmatrix}. \end{aligned}$$

It is easy to verify that the above estimate simplifies to (4) when all measurements have the same standard deviation σ .

3 Recursive Least Squares Estimation

Equation (5) is adequate when we have made all the measurements. More often, we obtain measurements sequentially and want to update our estimate with each new measurement. In this case, the matrix H needs to be augmented. We would have to recompute the estimate $\tilde{\mathbf{x}}$ according to (5) for every new measurement. This update can become very expensive. And the overall computation can become prohibitive as the number of measurements becomes large.

This section shows how to recursively compute the weighted least squares estimate. More specifically, suppose we have an estimate $\tilde{\mathbf{x}}_{k-1}$ after $k-1$ measurements, and obtain a new measurement \mathbf{y}_k . To be general, every measurement is now an m -vector with values yielded by, say, several measuring instruments. How can we update the estimate to $\tilde{\mathbf{x}}_k$ without solving equation (5)?

A linear recursive estimator can be written in the following form:

$$\begin{aligned}\mathbf{y}_k &= H_k \mathbf{x} + \boldsymbol{\nu}_k, \\ \tilde{\mathbf{x}}_k &= \tilde{\mathbf{x}}_{k-1} + K_k (\mathbf{y}_k - H_k \tilde{\mathbf{x}}_{k-1}).\end{aligned}\quad (6)$$

Here H_k is an $m \times n$ matrix, and K_k is $n \times m$ and referred to as the *estimator gain matrix*. We refer to $\mathbf{y}_k - H_k \tilde{\mathbf{x}}_{k-1}$ as the *correction term*. Namely, the new estimate $\tilde{\mathbf{x}}_k$ is modified from the previous estimate $\tilde{\mathbf{x}}_{k-1}$ with a correction via the gain vector. The measurement noise has zero mean, i.e., $E(\boldsymbol{\nu}_k) = \mathbf{0}$.

The current estimation error is

$$\begin{aligned}\boldsymbol{\epsilon}_k &= \mathbf{x} - \tilde{\mathbf{x}}_k \\ &= \mathbf{x} - \tilde{\mathbf{x}}_{k-1} - K_k (\mathbf{y}_k - H_k \tilde{\mathbf{x}}_{k-1}) \\ &= \boldsymbol{\epsilon}_{k-1} - K_k (H_k \mathbf{x} + \boldsymbol{\nu}_k - H_k \tilde{\mathbf{x}}_{k-1}) \\ &= \boldsymbol{\epsilon}_{k-1} - K_k H_k (\mathbf{x} - \tilde{\mathbf{x}}_{k-1}) - K_k \boldsymbol{\nu}_k \\ &= (I - K_k H_k) \boldsymbol{\epsilon}_{k-1} - K_k \boldsymbol{\nu}_k,\end{aligned}\quad (7)$$

where I is the $n \times n$ identity matrix. The mean of this error is then

$$E(\boldsymbol{\epsilon}_k) = (I - K_k H_k) E(\boldsymbol{\epsilon}_{k-1}) - K_k E(\boldsymbol{\nu}_k).$$

If $E(\boldsymbol{\nu}_k) = \mathbf{0}$ and $E(\boldsymbol{\epsilon}_{k-1}) = \mathbf{0}$, then $E(\boldsymbol{\epsilon}_k) = \mathbf{0}$. So if the measurement noise $\boldsymbol{\nu}_k$ has zero mean for all k , and the initial estimate of \mathbf{x} is set equal to its expected value, then $\tilde{\mathbf{x}}_k = \mathbf{x}_k$ for all k . With this property, the estimator (6) is called *unbiased*. The property holds regardless of the value of the gain vector K_k . It says that on the average the estimate $\tilde{\mathbf{x}}$ will be equal to the true value \mathbf{x} .

The key is to determine the optimal value of the gain vector K_k . The optimality criterion used by us is to *minimize the aggregated variance of the estimation errors at time k* :

$$\begin{aligned}J_k &= E(\|\mathbf{x} - \tilde{\mathbf{x}}_k\|^2) \\ &= E(\boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_k) \\ &= E(\text{Tr}(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^T)) \\ &= \text{Tr}(P_k),\end{aligned}\quad (8)$$

where Tr is the trace operator¹, and the $n \times n$ matrix $P_k = E(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^T)$ is the estimation-error covariance. Next, we obtain P_k via a substitution of (7):

$$\begin{aligned}P_k &= E\left(\left((I - K_k H_k) \boldsymbol{\epsilon}_{k-1} - K_k \boldsymbol{\nu}_k\right) \left((I - K_k H_k) \boldsymbol{\epsilon}_{k-1} - K_k \boldsymbol{\nu}_k\right)^T\right) \\ &= (I - K_k H_k) E(\boldsymbol{\epsilon}_{k-1} \boldsymbol{\epsilon}_{k-1}^T) (I - K_k H_k)^T - K_k E(\boldsymbol{\nu}_k \boldsymbol{\nu}_k^T) (I - K_k H_k)^T \\ &\quad - (I - K_k H_k) E(\boldsymbol{\epsilon}_{k-1} \boldsymbol{\nu}_k^T) K_k^T + K_k E(\boldsymbol{\nu}_k \boldsymbol{\nu}_k^T) K_k^T.\end{aligned}$$

The estimation error $\boldsymbol{\epsilon}_{k-1}$ at time $k-1$ is independent of the measurement noise $\boldsymbol{\nu}_k$ at time k , which implies that

$$\begin{aligned}E(\boldsymbol{\nu}_k \boldsymbol{\epsilon}_{k-1}^T) &= E(\boldsymbol{\nu}_k) E(\boldsymbol{\epsilon}_{k-1}^T) = 0, \\ E(\boldsymbol{\epsilon}_{k-1} \boldsymbol{\nu}_k^T) &= E(\boldsymbol{\epsilon}_{k-1}) E(\boldsymbol{\nu}_k^T) = 0.\end{aligned}$$

¹The trace of a matrix is the sum of its diagonal elements.

Given the definition of the $m \times m$ matrix $R_k = E(\boldsymbol{\nu}_k \boldsymbol{\nu}_k^T)$ as covariance of $\boldsymbol{\nu}_k$, the expression of P_k becomes

$$P_k = (I - K_k H_k) P_{k-1} (I - K_k H_k)^T + K_k R_k K_k^T. \quad (9)$$

Equation (9) is the recurrence for the covariance of the least squares estimation error. It is consistent with the intuition that as the measurement noise (R_k) increases, the uncertainty (P_k) increases. Note that P_k as a covariance matrix is positive definite.

What remains is to find the value of the gain vector K_k that minimizes the cost function given by (7). The mean of the estimation error is zero independent of the value of K_k already. Thus, the minimizing value of K_k will make the cost function consistently close to zero. We need to differentiate J_k with respect to K_k .²

Theorem 1 *Let C and X be matrices of the same dimension $r \times s$. Suppose C does not depend on X . Then the following holds:*

$$\frac{\partial \text{Tr}(C X^T)}{\partial X} = C, \quad (10)$$

$$\frac{\partial \text{Tr}(X C X^T)}{\partial X} = X C + X C^T. \quad (11)$$

A proof of the theorem is given in Appendix A. In the case that C is symmetric, $\frac{\partial}{\partial X} \text{Tr}(X C X^T) = 2X C$. With these facts in mind, we first substitute (9) into (8) and differentiate the resulting expression with respect to K_k :

$$\begin{aligned} \frac{\partial J_k}{\partial K_k} &= \frac{\partial}{\partial K_k} \text{Tr} \left(P_{k-1} - K_k H_k P_{k-1} - P_{k-1} H_k^T K_k^T + K_k (H_k P_{k-1} H_k^T) K_k^T \right) + \frac{\partial}{\partial K_k} \text{Tr}(K_k R_k K_k^T) \\ &= -2 \frac{\partial}{\partial K_k} \text{Tr}(P_{k-1} H_k^T K_k^T) + 2 K_k (H_k P_{k-1} H_k^T) + 2 K_k R_k \quad (\text{by (11)}) \\ &= -2 P_{k-1} H_k^T + 2 K_k H_k P_{k-1} H_k^T + 2 K_k R_k \quad (\text{by (10)}) \\ &= -2 P_{k-1} H_k^T + 2 K_k (H_k P_{k-1} H_k^T + R_k) \end{aligned}$$

In the second equation above, we also used that P_{k-1} is independent of K_k and that $K_k H_k P_{k-1}$ and $P_{k-1} H_k^T K_k^T$ are transposes of each other (since P_{k-1} is symmetric) so they have the same trace. Setting the partial derivative to zero, we solve for K_k :

$$K_k = P_{k-1} H_k^T (H_k P_{k-1} H_k^T + R_k)^{-1}. \quad (12)$$

Write $S_k = H_k P_{k-1} H_k^T + R_k$, so

$$K_k = P_{k-1} H_k^T S_k^{-1}. \quad (13)$$

Substitute the above for K_k into equation (9) for P_k . The operation followed by an expansion leads to a few steps of manipulation as follows:

$$P_k = (I - P_{k-1} H_k^T S_k^{-1} H_k) P_{k-1} (I - P_{k-1} H_k^T S_k^{-1} H_k)^T + P_{k-1} H_k^T S_k^{-1} R_k S_k^{-1} H_k P_{k-1}$$

²The derivative of a function f with respect to a matrix $A = (a_{ij})$ is a matrix $\partial f / \partial A = (\partial f / \partial a_{ij})$.

$$\begin{aligned}
&= P_{k-1} - P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} - P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} + \\
&\quad P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} \underline{H_k P_{k-1} H_k^T} S_k^{-1} H_k P_{k-1} + P_{k-1}H_k^T S_k^{-1} \underline{R_k} S_k^{-1} H_k P_{k-1} \\
&= P_{k-1} - P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} - P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} + P_{k-1}H_k^T S_k^{-1} S_k S_k^{-1} H_k P_{k-1} \\
&\quad \text{(after merging the underlined terms into } S_k) \\
&= P_{k-1} - 2P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} + P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} \\
&= P_{k-1} - P_{k-1}H_k^T S_k^{-1} H_k P_{k-1} \tag{14} \\
&= P_{k-1} - K_k H_k P_{k-1} \quad \text{(by (13))} \\
&= (I - K_k H_k) P_{k-1}. \tag{15}
\end{aligned}$$

Note that in the above P_k is symmetric as a covariance matrix, and so is S_k .

We take the inverses of both sides of equation (14) and plug into the expression for S_k :

$$P_k^{-1} = \left(\underbrace{P_{k-1}}_A - \underbrace{P_{k-1}H_k^T}_B \underbrace{(H_k P_{k-1} H_k^T + R_k)}_D \underbrace{H_k P_{k-1}}_C \right)^{-1}.$$

Next, we apply the matrix inversion lemma:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

and rewrite P_k^{-1} as follows:

$$\begin{aligned}
P_k^{-1} &= P_{k-1}^{-1} + P_{k-1}^{-1} P_{k-1} H_k^T \left((H_k P_{k-1} H_k^T + R_k) - H_k P_{k-1} P_{k-1}^{-1} (P_{k-1} H_k^T) \right)^{-1} H_k P_{k-1} P_{k-1}^{-1} \\
&= P_{k-1}^{-1} + H_k^T R_k^{-1} H_k. \tag{16}
\end{aligned}$$

The above yields an alternative expression for the covariance matrix:

$$P_k = (P_{k-1}^{-1} + H_k^T R_k^{-1} H_k)^{-1}. \tag{17}$$

This expression is more complicated than (15) since it requires three matrix inversions. Nevertheless, it has computational advantages in certain situations in practice [1, pp.156–158].

We can also derive an alternate form for the covariance P_k as follows. Start with a multiplication of the right of (12) with $P_k P_k^{-1}$. Then, substitute (16) for P_k^{-1} into the resulting expression. Multiply the $P_{k-1} H_k$ factor inside the parenthesized factor on its left, and extract $H_k^T R_k^{-1}$ out of the parentheses. The last two parenthesized factors will cancel each other, yielding

$$K_k = P_k H_k^T R_k^{-1}. \tag{18}$$

4 The Estimation Algorithm

The algorithm for recursive least squares estimation is summarized as follows.

1. Initialize the estimator:

$$\begin{aligned}
\tilde{\mathbf{x}}_0 &= E(\mathbf{x}), \\
P_0 &= E((\mathbf{x} - \tilde{\mathbf{x}}_0)(\mathbf{x} - \tilde{\mathbf{x}}_0)^T).
\end{aligned}$$

In the case of no prior knowledge about \mathbf{x} , simply let $P_0 = \infty I$. In the case of perfect prior knowledge, let $P_0 = 0$.

2. Iterate the follow two steps.

- (a) Obtain a new measurement \mathbf{y}_k , assuming that it is given by the equation

$$\mathbf{y}_k = H_k \mathbf{x} + \boldsymbol{\nu}_k,$$

where the noise $\boldsymbol{\nu}_k$ has zero mean and covariance R_k . The measurement noise at each time step k is independent. So,

$$E(\boldsymbol{\nu}_i \boldsymbol{\nu}_j^T) = \begin{cases} 0, & \text{if } i \neq j, \\ R_j, & \text{if } i = j. \end{cases}$$

Essentially, we assume white measurement noise.

- (b) Update the estimate $\tilde{\mathbf{x}}$ and the covariance of the estimation error sequentially according to (12), (6), (15), which are re-listed below:

$$K_k = P_{k-1} H_k^T (H_k P_{k-1} H_k^T + R_k)^{-1}, \quad (19)$$

$$P_k = (I - K_k H_k) P_{k-1}, \quad (20)$$

$$\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{k-1} + K_k (\mathbf{y}_k - H_k \tilde{\mathbf{x}}_{k-1}), \quad (21)$$

or according to (17), (18), and (21):

$$P_k = (P_{k-1}^{-1} + H_k^T R_k^{-1} H_k)^{-1},$$

$$K_k = P_k H_k^T R_k^{-1},$$

$$\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{k-1} + K_k (\mathbf{y}_k - H_k \tilde{\mathbf{x}}_{k-1}).$$

Note that (21) and (20) can switch their order in one round of update.

EXAMPLE 3. We revisit the resistance estimation problem presented in Examples 1 and 2. Now, we want to iteratively improve our estimate of the resistance x . At the k th sampling, our measurement is

$$\begin{aligned} y_k &= H_k x + \nu_k = x + \nu_k, \\ R_k &= E(\nu_k^2). \end{aligned}$$

Here, the measurement vector H_k is a scalar 1. Furthermore, we suppose that each measurement has the same covariance so R_k is a constant written as R .

Before the first measurement, we have some idea about the resistance x . This becomes our initial estimate. Also, we have some uncertainty about this initial estimate, which becomes our initial covariance. Together we have

$$\begin{aligned} \tilde{x}_0 &= E(x), \\ P_0 &= E((x - \tilde{x}_0)^2). \end{aligned}$$

If we have no idea about the resistance, set $P_0 = \infty$. If we are certain about the resistance value, set $P_0 = 0$. (Of course, then there would be no need to take measurements.)

After the first measurement ($k=1$), we update the estimate and the error covariance according to equations (19)–(20) as follows:

$$\begin{aligned} K_1 &= \frac{P_0}{P_0 + R}, \\ \tilde{x}_1 &= \tilde{x}_0 + \frac{P_0}{P_0 + R} (y_1 - \tilde{x}_0), \\ P_1 &= \left(1 - \frac{P_0}{P_0 + R}\right) P_0 = \frac{P_0 R}{P_0 + R}. \end{aligned}$$

After the second measurement, the estimates become

$$\begin{aligned} K_2 &= \frac{P_1}{P_1 + R} = \frac{P_0}{2P_0 + R}, \\ \tilde{x}_2 &= \tilde{x}_1 + \frac{P_1}{P_1 + R}(y_2 - \tilde{x}_1) \\ &= \frac{P_0 + R}{2P_0 + R}\tilde{x}_1 + \frac{P_0}{2P_0 + R}y_2, \\ P_2 &= \frac{P_1 R}{P_1 + R} = \frac{P_0 R}{2P_0 + R}. \end{aligned}$$

By induction, we can show that

$$\begin{aligned} K_k &= \frac{P_0}{kP_0 + R}, \\ \tilde{x}_k &= \frac{(k-1)P_0 + R}{kP_0 + R}\tilde{x}_{k-1} + \frac{P_0}{kP_0 + R}y_k, \\ P_k &= \frac{P_0 R}{kP_0 + R}. \end{aligned}$$

Note that if x is known perfectly *a priori*, then $P_0 = 0$, which implies that $K_k = 0$ and $\tilde{x}_k = \tilde{x}_0$, for all k . The optimal estimate of x is independent of any measurements that are obtained. At the opposite end of the spectrum, if x is completely unknown *a priori*, then $P_0 = \infty$. The above equation for \tilde{x}_k becomes,

$$\begin{aligned} \tilde{x}_k &= \lim_{P_0 \rightarrow \infty} \frac{(k-1)P_0 + R}{kP_0 + R}\tilde{x}_{k-1} + \frac{P_0}{kP_0 + R}y_k \\ &= \frac{k-1}{k}\tilde{x}_{k-1} + \frac{1}{k}y_k \\ &= \frac{1}{k}\left((k-1)\tilde{x}_{k-1} + y_k\right). \end{aligned}$$

The right hand side of the last equation above is just the running average $\bar{y}_k = \frac{1}{k} \sum_{j=1}^k y_j$ of the measurements. To see this, we first have

$$\begin{aligned} \sum_{j=1}^k y_j &= \sum_{j=1}^{k-1} y_j + y_k \\ &= (k-1) \left(\frac{1}{k-1} \sum_{j=1}^{k-1} y_j \right) + y_k \\ &= (k-1)\bar{y}_{k-1} + y_k. \end{aligned}$$

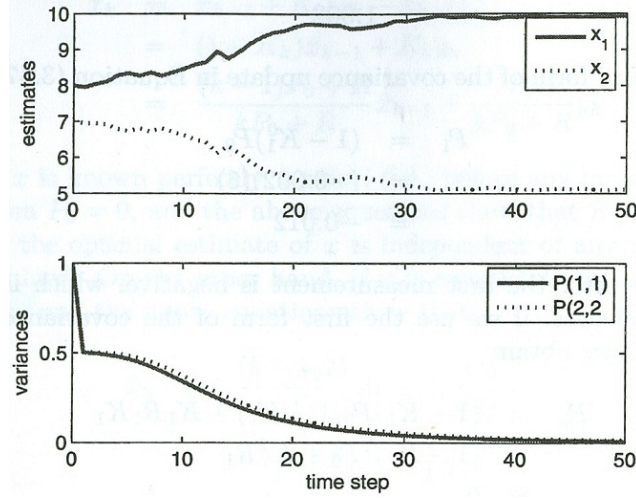
Since $\tilde{x}_1 = \bar{y}_1$, the recurrences for \tilde{x}_k and \bar{y}_k are the same. Hence $\tilde{x}_k = \bar{y}_k$ for all k .

EXAMPLE 4. Suppose that a tank contains a concentration x_1 of chemical 1, and a concentration x_2 of chemical 2. We have an instrument to detect the combined concentration $x_1 + x_2$ of the two chemicals but not able to tell the values of x_1 and x_2 . Chemical 2 leaks from the tank so that its concentration decreases by 1% from one measurement to the next. The measurement equation is given as

$$y_k = x_1 + 0.99^{k-1}x_2 + \nu_k,$$

where $H_k = (1, 0.99^{k-1})^T$, and ν_k is a random variable with zero mean and a variance $R = 0.01$.

Let the real values be $\tilde{x} = (x_1, x_2)^T = (10, 5)^T$. Suppose the initial estimates are $\tilde{x}_1 = 8$ and $\tilde{x}_2 = 7$ with P_0 equal to the identity matrix. We apply the recursive least squares algorithm. The next figure³ shows the evolutions of the estimates \tilde{x}_1 and \tilde{x}_2 , along with those of the variance of the estimation errors. It can be seen that after a couple dozen measurements, the estimates are getting very close to the true values 10 and 5. The variances of the estimation errors asymptotically approach zero. This means that we have increasingly more confidence in the estimates with more measurements obtained.



A Proof of Theorem 1

Proof

Denote $C = (c_{ij})$, $X = (x_{ij})$, and $CX^T = (d_{ij})$. The trace of CX^T is

$$\begin{aligned} \text{Tr}(CX^T) &= \sum_{t=1}^r d_{tt} \\ &= \sum_{t=1}^r \sum_{k=1}^s c_{tk} x_{tk}. \end{aligned}$$

From the above, we easily obtain its partial derivatives with respect to the entries of X :

$$\frac{\partial}{\partial x_{ij}} \text{Tr}(CX^T) = c_{ij}.$$

This establishes (10).

To prove (11), we have

$$\begin{aligned} \frac{\partial}{\partial X} \text{Tr}(XCX^T) &= \frac{\partial}{\partial X} \text{Tr}(XCY^T) \Big|_{Y=X} + \frac{\partial}{\partial X} \text{Tr}(YCX^T) \Big|_{Y=X} \\ &= \frac{\partial}{\partial X} \text{Tr}(YC^T X^T) \Big|_{Y=X} + YC \Big|_{Y=X} \quad (\text{by (10)}) \end{aligned}$$

³Figure 3.1, p. 92 of [1].

$$\begin{aligned} &= YC^T \Big|_{Y=X} + XC \\ &= XC^T + XC. \end{aligned}$$

□

References

- [1] D. Simon. *Optimal State Estimations*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.