Danilo P. Mandic, Sithan Kanna
and Anthony G. Constantinides

On the Intrinsic Relationship Between the Least Mean Square and Kalman Filters

The Kalman filter and the least mean square (LMS) adaptive filter are two of the most popular adaptive estimation algorithms that are often used interchangeably in a number of statistical signal processing applications. They are typically treated as separate entities, with the former as a realization of the optimal Bayesian estimator and the latter as a recursive solution to the optimal Wiener filtering problem. In this lecture note, we consider a system identification framework within which we develop a joint perspective on Kalman filtering and LMS-type algorithms, achieved through analyzing the degrees of freedom necessary for optimal stochastic gradient descent adaptation. This approach permits the introduction of Kalman filters without any notion of Bayesian statistics, which may be beneficial for many communities that do not rely on Bayesian methods [1], [2].

There are several and not immediately patent aspects of common thinking between gradient descent and recursive state-space estimators. Because of their nonobvious or awkward nature, these are often overlooked. Hopefully the framework presented in this article, with the seamless transition between LMS and Kalman filters, will provide a straightforward and unifying platform for understanding the geometry of learning and optimal parameter selection in these approaches. In addition, the material may be useful in lecture courses in statistical signal processing, or indeed, as interesting reading for the intellectually curious and generally knowledgeable reader.

Digital Object Identifier 10.1109/MSP.2015.2461733
Date of publication: 13 October 2015

NOTATION

Lowercase letters are used to denote scalars, e.g., a; boldface letters for vectors, $\mathbf{a}$; and boldface uppercase letters for matrices, $\mathbf{A}$. Vectors and matrices are respectively of dimensions $M \times 1$ and $M \times M$. The symbol $(\cdot)^T$ is used for vector and matrix transposition and the subscript k for discrete time index. Symbol $E\{\cdot\}$ represents the statistical expectation operator, $\mathrm{tr}\{\cdot\}$ is the matrix trace operator, and $\|\cdot\|^2$ the l_2 norm.

PROBLEM FORMULATION

Consider a generic system identification setting

$$d_k = \mathbf{x}_k^T \mathbf{w}_k^o + n_k, \tag{1}$$

where the aim is to estimate the unknown true system parameter vector, $\mathbf{w}_k^o$ (optimal weight vector), which characterizes the system in (1) from observations, d_k, corrupted by observation noise, n_k. This parameter vector can be fixed, i.e., $\mathbf{w}_k^o = \mathbf{w}^o$, or time varying as in (1), while $\mathbf{x}_k$ designates a zero-mean input vector and n_k is a zero-mean white Gaussian process with variance $\sigma_n^2 = E\{n_k^2\}$. For simplicity, we assume that all signals are real valued.

To assist a joint discussion of state-space and regression-type models Table 1 lists the terms commonly used across different communities for the variables in the system identification paradigm in (1).

We first start the discussion with a deterministic and time-invariant optimal weight vector, $\mathbf{w}_k^{\mathrm{o}} = \mathbf{w}^{\mathrm{o}}$, and build up to the general case of a stochastic and time-varying system to give the general Kalman filter.

PERFORMANCE EVALUATION CRITERIA

Consider observations from an unknown deterministic system

$$d_k = \mathbf{x}_k^{\mathrm{T}} \mathbf{w}^{\mathrm{o}} + n_k. \tag{2}$$

We desire to estimate the true parameter vector $\mathbf{w}^{\mathrm{o}}$ recursively, based on the existing weight vector estimate $\mathbf{w}_{k-1}$ and the observed and input signals, i.e., $\hat{\mathbf{w}}^{\mathrm{o}} = \mathbf{w}_k = f(\mathbf{w}_{k-1}, d_k, \mathbf{x}_k)$. Notice that $\mathbf{w}_{k-1}, d_k, \mathbf{x}_k$ are related through the output error

$$e_k = d_k - \mathbf{x}_k^{\mathrm{T}} \mathbf{w}_{k-1}. \tag{3}$$

Performance of statistical learning algorithms is typically evaluated based on the mean square error (MSE) criterion, which is defined as the output error power and is given by

$$\mathrm{MSE} = \xi_k \overset{\mathrm{def}}{=} E\{e_k^2\}. \tag{4}$$

Since our goal is to estimate the true system parameters, it is natural to also consider the weight error vector

$$\tilde{\mathbf{w}}_k \overset{\mathrm{def}}{=} \mathbf{w}^{\mathrm{o}} - \mathbf{w}_k, \tag{5}$$

and its contribution to the output error, given by

$$e_k = \mathbf{x}_k^{\mathrm{T}} \tilde{\mathbf{w}}_{k-1} + n_k. \tag{6}$$

[TABLE 1] THE TERMINOLOGY USED IN DIFFERENT COMMUNITIES.

AREA	d_k	$\mathbf{x}_k$	$\mathbf{w}_k^{\mathrm{O}}$
ADAPTIVE FILTERING	DESIRED SIGNAL	INPUT REGRESSOR	TRUE/OPTIMAL WEIGHTS
KALMAN FILTERING	OBSERVATION	MEASUREMENT	STATE VECTOR
MACHINE LEARNING	TARGET	FEATURES	HYPOTHESIS PARAMETERS

Without loss of generality, here we treat $\mathbf{x}_k$ as a deterministic process, although in adaptive filtering convention it is assumed to be a zero-mean stochastic process with covariance matrix $\mathbf{R} = E\{\mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}\}$. Our assumption conforms with the Kalman filtering literature, where the vector $\mathbf{x}_k$ is often deterministic (and sometimes even time invariant). Replacing the output error from (6) into (4) gives

$$\begin{aligned} \xi_k &= E\{(\mathbf{x}_k^{\mathrm{T}} \tilde{\mathbf{w}}_{k-1} + n_k)^2\} \\ &= \mathbf{x}_k^{\mathrm{T}} \mathbf{P}_{k-1} \mathbf{x}_k + \sigma_n^2 && (7a) \\ &\overset{\mathrm{def}}{=} \xi_{\mathrm{ex},k} + \xi_{\mathrm{min}}, && (7b) \end{aligned}$$

where $\mathbf{P}_{k-1} \overset{\mathrm{def}}{=} E\{\tilde{\mathbf{w}}_{k-1} \tilde{\mathbf{w}}_{k-1}^{\mathrm{T}}\}$ is the symmetric and positive semidefinite weight error covariance matrix, and the noise process n_k is assumed to be statistically independent from all other variables. Therefore, for every recursion step, k, the corresponding MSE denoted by ξ_k comprises two terms: 1) the time-varying excess MSE (EMSE), $\xi_{\mathrm{ex},k}$, which reflects the misalignment between the true and estimated weights (function of the performance of the estimator), and 2) the observation noise power, $\xi_{\mathrm{min}} = \sigma_n^2$, which represents the minimum achievable MSE (for $\mathbf{w}_k = \mathbf{w}^{\circ}$) and is independent of the performance of the estimator.

Our goal is to evaluate the performance of a learning algorithm in identifying the true system parameters, $\mathbf{w}^{\circ}$, and a more insightful measure of how closely the estimated weights, $\mathbf{w}_k$, have approached the true weights, $\mathbf{w}^{\circ}$, is the mean square deviation (MSD), which represents the power of the weight error vector and is given by

$$\begin{aligned} \mathrm{MSD} = J_k \overset{\mathrm{def}}{=} E\{\|\tilde{\mathbf{w}}_k\|^2\} &= E\{\tilde{\mathbf{w}}_k^{\mathrm{T}} \tilde{\mathbf{w}}_k\} \\ &= \mathrm{tr}\{\mathbf{P}_k\}. && (8) \end{aligned}$$

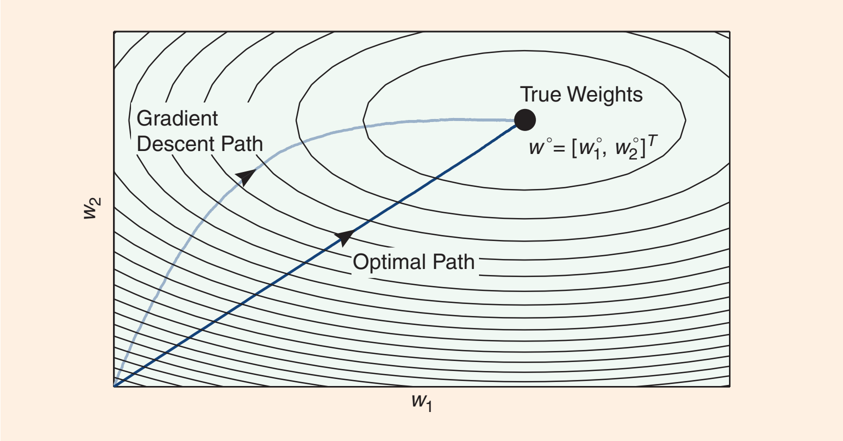
Observe that the MSD is related to the MSE in (7a) through the weight error covariance matrix, $\mathbf{P}_k = E\{\tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^{\mathrm{T}}\}$, and thus minimizing MSD also corresponds to minimizing MSE.

OPTIMAL LEARNING GAIN FOR STOCHASTIC GRADIENT ALGORITHMS

The LMS algorithm employs stochastic gradient descent to approximately minimize the MSE in (4) through a recursive estimation of the optimal weight vector, $\mathbf{w}^{\circ}$ in (2), in the form $\mathbf{w}_k = \mathbf{w}_{k-1} - \mu_k \nabla_{\mathbf{w}} E\{e_k^2\}$. Based on the instantaneous estimate $E\{e_k^2\} \approx e_k^2$, the LMS solution is then given by [3]

$$\begin{aligned} \text{LMS:} \quad \mathbf{w}_k &= \mathbf{w}_{k-1} + \Delta \mathbf{w}_k \\ &= \mathbf{w}_{k-1} + \mu_k \mathbf{x}_k e_k. && (9) \end{aligned}$$

The parameter μ_k is a possibly time-varying positive step-size that controls the magnitude of the adaptation steps the algorithm takes; for fixed system parameters this can be visualized as a trajectory along the error surface—the MSE plot evaluated against the weight vector, $\xi_k(\mathbf{w})$. Notice that the weight update $\Delta \mathbf{w}_k = \mu_k \mathbf{x}_k e_k$ has the same direction as the input signal vector, $\mathbf{x}_k$, which makes the LMS sensitive to outliers and noise in data. Figure 1 illustrates the geometry of learning of gradient descent approaches for correlated data (elliptical contours of the error surface)—gradient descent performs locally optimal steps but has no means to follow the globally optimal shortest path to the solution, $\mathbf{w}^{\circ}$. It is therefore necessary to control both the direction and magnitude of adaptation steps for an algorithm to follow the shortest, optimal path to the global minimum of error surface, $\xi(\mathbf{w}^{\circ})$.



[FIG1] Mean trajectories of an ensemble of noisy single-realization gradient descent paths for correlated data. The LMS path, produced based on (9), is locally optimal but globally slower converging than the optimal path.

The first step toward Kalman filters is to introduce more degrees of freedom by replacing the scalar step-size, μ_k, with a positive definite learning gain matrix, $\mathbf{G}_k$, so as to control both the magnitude and direction of the gradient descent adaptation, and follow the optimal path in Figure 1. In this way, the weight update recursion in (9) now generalizes to

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \mathbf{G}_k \mathbf{x}_k e_k. \qquad (10)$$

Unlike standard gradient-adaptive step-size approaches that minimize the MSE via $\partial \xi_k / \partial \mu_k$ [4], [5], our aim is to introduce an optimal step-size (and learning gain) into the LMS based on the direct minimization of the MSD in (8). For convenience, we consider a general recursive weight estimator

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \mathbf{g}_k e_k, \qquad (11)$$

which represents both (9) and (10), where the gain vector

$$\mathbf{g}_k \overset{\mathrm{def}}{=} \begin{cases} \mu_k \mathbf{x}_k, & \text{for the conventional LMS in (9),} \\ \mathbf{G}_k \mathbf{x}_k, & \text{for a general LMS in (10).} \end{cases} \qquad (12)$$

To minimize the MSD, given by $J_k = E\{\|\tilde{\mathbf{w}}_k\|^2\} = \mathrm{tr}\{\mathbf{P}_k\}$, we first establish the weight error vector recursion for the general LMS by subtracting $\mathbf{w}^{\circ}$ from both sides of (11) and replacing the output error with $e_k = \mathbf{x}_k^{\mathrm{T}} \tilde{\mathbf{w}}_{k-1} + n_k$, to give

$$\tilde{\mathbf{w}}_k = \tilde{\mathbf{w}}_{k-1} - \mathbf{g}_k \mathbf{x}_k^{\mathrm{T}} \tilde{\mathbf{w}}_{k-1} - \mathbf{g}_k n_k. \qquad (13)$$

The recursion for the weight error covariance matrix, $\mathbf{P}_k$, is then established upon postmultiplying both sides of (13) by their respective transposes and applying the statistical expectation operator $E\{\cdot\}$ to both sides, to yield

$$\begin{aligned}\mathbf{P}_k &= E\{\tilde{\boldsymbol{w}}_k\tilde{\boldsymbol{w}}_k^{\mathrm{T}}\}\\ &= \mathbf{P}_{k-1}-(\mathbf{P}_{k-1}\boldsymbol{x}_k\mathbf{g}_k^{\mathrm{T}}+\mathbf{g}_k\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1})\\ &\quad+\mathbf{g}_k\mathbf{g}_k^{\mathrm{T}}(\boldsymbol{x}_k^{\mathrm{T}}P_{k-1}\boldsymbol{x}_k+\sigma_n^2). \qquad (14)\end{aligned}$$

Using the well-known matrix trace identities, $\mathrm{tr}\{\mathbf{P}_{k-1}\boldsymbol{x}_k\mathbf{g}_k^{\mathrm{T}}\}=\mathrm{tr}\{\mathbf{g}_k\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\}$ $=\mathbf{g}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k$ and $\mathrm{tr}\{\mathbf{g}_k\mathbf{g}_k^{\mathrm{T}}\}=\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k=\|\mathbf{g}_k\|^2$, the MSD evolution, $J_k=\mathrm{tr}\{\mathbf{P}_k\}$, is obtained as

$$\begin{aligned}J_k &= J_{k-1}-2\mathbf{g}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k\\ &\quad+\|\mathbf{g}_k\|^2(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2). \qquad (15)\end{aligned}$$

OPTIMAL SCALAR STEP-SIZE FOR LMS

The standard optimal step-size approach to the LMS aims at achieving $e_{k+1|k}=d_k-\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_k=0$, where the a posteriori error, $e_{k+1|k}$, is obtained using the updated weight vector, $\boldsymbol{w}_k$, and the current input, $\boldsymbol{x}_k$. The solution is known as the normalized LMS (NLMS), given by (for more details, see [6])

$$\text{NLMS:}\quad \boldsymbol{w}_k=\boldsymbol{w}_{k-1}+\frac{1}{\|\boldsymbol{x}_k\|^2}\boldsymbol{x}_k e_k. \qquad (16)$$

The effective LMS-type step-size, $\mu_k=1/\|\boldsymbol{x}_k\|^2$, is now time varying and data adaptive. In practice, to stabilize the algorithm a small positive step-size ρ_k can be employed, to give $\mu_k=\rho_k/\|\boldsymbol{x}_k\|^2$. The NLMS is therefore conformal with the LMS, whereby the input vector, $\boldsymbol{x}_k$, is normalized by its norm, $\|\boldsymbol{x}_k\|^2$ (input signal power).

To find the optimal scalar step-size for the LMS in (9), which minimizes the MSD, we shall first substitute the gain $\mathbf{g}_k=\mu_k\boldsymbol{x}_k$ into (15), to give the MSD recursion

$$\begin{aligned}J_k &= J_{k-1}-2\mu_k\underbrace{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}_{\xi_{ex,k}}\\ &\quad+\mu_k^2\|\boldsymbol{x}_k\|^2\underbrace{(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2)}_{\xi_k}. \qquad (17)\end{aligned}$$

The optimal step-size, which minimizes MSD, is then obtained by solving for μ_k in (17) via $\partial J_k/\partial\mu_k=0$, to yield [7]

$$\begin{aligned}\mu_k &= \frac{1}{\|\boldsymbol{x}_k\|^2}\frac{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}{(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2)}\\ &= \underbrace{\frac{1}{\|\boldsymbol{x}_k\|^2}}_{\text{normalization}}\ \underbrace{\frac{\xi_{\mathrm{ex},k}}{\xi_k}}_{\text{correction}}. \qquad (18)\end{aligned}$$

REMARK 1

In addition to the NLMS-type normalization factor, $1/\|\boldsymbol{x}_k\|^2$, the optimal LMS step-size in (18) includes the correction term, $\xi_{\mathrm{ex},k}/\xi_k<1$, a ratio of the EMSE, $\xi_{\mathrm{ex,\,k,}}$ to the overall MSE, ξ_k. A large deviation from the true system weights causes a large $\xi_{\mathrm{ex},\,k}/\xi_k$ and fast weight adaptation (cf. slow adaptation for a small $\xi_{\mathrm{ex},\,k}/\xi_k$). This also justifies the use of a small step-size, ρ_k, in practical NLMS algorithms, such as that in "Variants of the LMS."

FROM LMS TO KALMAN FILTER

The optimal LMS step-size in (18) aims to minimize the MSD at every time instant, however, it only controls the magnitude of gradient descent steps (see Figure 1). To find the optimal learning gain that controls simultaneously both the magnitude and direction of the gradient descent in (10), we start again from the MSD recursion [restated from (15)]

$$\begin{aligned}J_k &= J_{k-1}-2\mathbf{g}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k\\ &\quad+\|\mathbf{g}_k\|^2(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2).\end{aligned}$$

The optimal learning gain vector, $\mathbf{g}_k$, is then obtained by solving the above MSD for $\mathbf{g}_k$, via $\partial J_k/\partial g_k=0$, to give

$$\begin{aligned}\mathbf{g}_k &= \frac{\mathbf{P}_{k-1}}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2}\boldsymbol{x}_k=\frac{\mathbf{P}_{k-1}}{\xi_k}\boldsymbol{x}_k\\ &= \mathbf{G}_k\boldsymbol{x}_k. \qquad (19)\end{aligned}$$

This optimal gain vector is precisely the Kalman gain [8], while the gain matrix, $\mathbf{G}_k$, represents a ratio between the weight error covariance, $\mathbf{P}_{k-1}$, and the MSE, ξ_k. A substitution into the update for $\mathbf{P}_k$ in (14) yields a Kalman filter that estimates the time-invariant and deterministic weights, $\boldsymbol{w}^{\mathrm{o}}$, as outlined in Algorithm 1.

REMARK 2

For $\sigma_n^2=1$, the Kalman filtering equations in Algorithm 1 are identical to the recursive least squares (RLS) algorithm. In this way, this lecture note complements the classic article by Sayed and Kailath [9] that establishes a relationship between the RLS and the Kalman filter.

SCALAR COVARIANCE UPDATE

An additional insight into our joint perspective on Kalman and LMS algorithms is provided for independent and identically distributed system weight error vectors, whereby the diagonal weight error

VARIANTS OF THE LMS

To illustrate the generality of our results, consider the NLMS and the regularized NLMS (also known as $\varepsilon-$NLMS), given by

$$\text{NLMS:}\quad \boldsymbol{w}_k=\boldsymbol{w}_{k-1}-\rho_k\frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|^2}e_k, \qquad \text{(S1)}$$

$$\varepsilon-\text{NLMS:}\quad \boldsymbol{w}_k=\boldsymbol{w}_{k-1}+\frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|^2+\varepsilon_k}e_k, \qquad \text{(S2)}$$

where ρ_k is a step-size and ε_k a regularization factor. Based on (17) and (18), the optimal values for ρ_k and ε_k can be found as

$$\rho_k=\frac{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2},\qquad \varepsilon_k\frac{\|\boldsymbol{x}_k\|^2\sigma_n^2}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}. \qquad \text{(S3)}$$

Upon substituting ρ_k and ε_k from (S3) into their respective weight update recursions in (S1) and (S2), we arrive at

$$\boldsymbol{w}_k=\boldsymbol{w}_{k-1}+\frac{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}{(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k+\sigma_n^2)}\frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|^2}e_k, \qquad \text{(S4)}$$

for both the NLMS and $\varepsilon-$NLMS, which is identical to the LMS with the optimal step-size in (18). Therefore, the minimization of the mean square deviation with respect to the parameter: 1) μ_k in the LMS, 2) ρ_k in the NLMS, and 3) ε_k in the ε-NLMS, yields exactly the same algorithm, which is intimately related to the Kalman filter, as shown in Table 2 and indicated by the expression for the Kalman gain, $\mathbf{g}_k$.

Algorithm 1: The Kalman filter for deterministic states.

```
At each time instant $k > 0$, based on measurements $\{d_k, \boldsymbol{x}_k\}$
1) Compute the optimal learning gain (Kalman gain):
    $\mathrm{g}_k = \mathrm{P}_{k-1}\boldsymbol{x}_k/(\mathbf{x}_k^{\mathrm{T}}\mathrm{P}_{k-1}\boldsymbol{x}_k + \sigma_n^2)$
2) Update the weight vector estimate:
    $\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \mathrm{g}_k(d_k - \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_{k-1})$
3) Update the weight error covariance matrix:
    $\mathrm{P}_k = \mathrm{P}_{k-1} - \mathrm{g}_k\boldsymbol{x}_k^{\mathrm{T}}\mathrm{P}_{k-1}$
```

covariance matrix is given by $\mathrm{P}_{k-1} = \sigma_{P,k-1}^2\mathbf{I}$, while the Kalman gain, g_k, in (19) now becomes

$$\mathrm{g}_k = \frac{\sigma_{P,k-1}^2}{\sigma_{P,k-1}^2\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k + \sigma_n^2}\boldsymbol{x}_k = \frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|^2 + \varepsilon_k}, \tag{20}$$

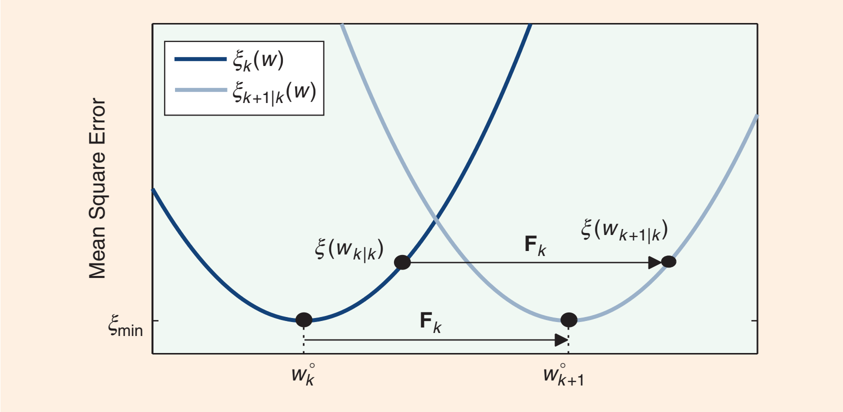
where $\varepsilon_k \overset{\text{def}}{=} \sigma_n^2/\sigma_{P,k-1}^2$ denotes the regularization parameter and $\sigma_{P,k-1}^2$ is the estimated weight error vector variance.

REMARK 3

A physical interpretation of the regularization parameter, ε_k, is that it models our confidence level in the current weight estimate, $\boldsymbol{w}_k$, via a ratio of the algorithm-independent minimum MSE, $\xi_{\min} = \sigma_n^2$, and the algorithm-specific weight error variance, $\sigma_{P,k-1}^2$. The more confident we are in current weight estimates, the greater the value of ε_k and the smaller the magnitude of the weight update, $\Delta\boldsymbol{w}_k = \mathrm{g}_k e_k$.

Algorithm 2: A hybrid Kalman-LMS algorithm.

```
At each time instant $k > 0$, based on measurements $\{d_k, \boldsymbol{x}_k\}$
1) Compute the confidence level (regularisation parameter):
    $\varepsilon_k = \sigma_n^2/\sigma_{P,k-1}^2$
2) Update the weight vector estimate:
    $\boldsymbol{w}_k = \boldsymbol{w}_{k-1} + \frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|^2 + \varepsilon_k}(d_k - \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_{k-1})$
3) Update the weight error variance:
    $\sigma_{P,k}^2 = \sigma_{P,k-1}^2 - \frac{\|\boldsymbol{x}_k\|^2}{M(\|\boldsymbol{x}_k\|^2 + \varepsilon_k)}\sigma_{P,k-1}^2$
```



[FIG2] The time-varying state transition in (22a) results in a time-varying MSE surface. For clarity, the figure considers a scalar case without state noise. Within the Kalman filter, the prediction step in (23b) preserves the relative position of $\boldsymbol{w}_{k+1|k}$ with respect to the evolved true state, $\boldsymbol{w}_{k+1}^{\circ}$.

To complete the derivation, since $\mathrm{P}_k = \sigma_{P,k}^2\mathbf{I}$ and $\mathrm{tr}\{\mathrm{P}_k\} = M\sigma_{P,k}^2$, the MSD recursion in (15) now becomes

$$\sigma_{P,k}^2 = \sigma_{P,k-1}^2 - \frac{\|\boldsymbol{x}_k\|^2}{M(\|\boldsymbol{x}_k\|^2 + \varepsilon_k)}\sigma_{P,k-1}^2. \tag{21}$$

The resulting hybrid "Kalman-LMS" algorithm is given in Algorithm 2.

REMARK 4

The form of the LMS algorithm outlined in Algorithm 2 is identical to the class of generalized normalized gradient descent (GNGD) algorithms in [5] and [10], which update the regularization parameter, ε_k, using stochastic gradient descent. More recently, Algorithm 2 was derived independently in [11] as an approximate probabilistic filter for linear Gaussian data and is referred to as the *probabilistic LMS*.

FROM OPTIMAL LMS TO GENERAL KALMAN FILTER

To complete the joint perspective on the LMS and Kalman filters, we now consider a general case of a time-varying and stochastic weight vector $\boldsymbol{w}_k^{\mathrm{o}}$ in (1), to give

$$\boldsymbol{w}_{k+1}^{\mathrm{o}} = \mathrm{F}_k\boldsymbol{w}_k^{\mathrm{o}} + \boldsymbol{q}_k, \quad \boldsymbol{q}_k \sim \mathcal{N}(0, \boldsymbol{Q}_s), \tag{22a}$$

$$d_k = \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_k^{\mathrm{o}} + n_k, \quad n_k \sim \mathcal{N}(0, \sigma_n^2). \tag{22b}$$

The evolution of the true weight vector $\boldsymbol{w}_k^{\mathrm{o}}$ is governed by a known state transition matrix, F_k, while the uncertainty in the state transition model is represented by a temporally white state noise vector, q_k, with covariance $\mathrm{Q}_s = E\{\mathrm{q}_k\mathrm{q}_k^{\mathrm{T}}\}$, which is uncorrelated with observation noise n_k. The optimal weight vector evolution in (22a) requires both the update of the current state estimate, $\boldsymbol{w}_{k|k}$, in an LMS-like fashion and the prediction of the next state, $\boldsymbol{w}_{k+1|k}$, as below

$$\boldsymbol{w}_{k|k} = \boldsymbol{w}_{k|k-1} + \mathrm{g}_k(d_k - \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_{k|k-1}), \tag{23a}$$

$$\boldsymbol{w}_{k+1|k} = \mathrm{F}_k\boldsymbol{w}_{k|k}, \tag{23b}$$

where g_k in (23a) is the Kalman gain. Figure 2 illustrates that, unlike the standard LMS or deterministic Kalman filter in Algorithm 1, the general Kalman filter in (23a) and (23b) employs its prediction step in (23b) to track the time-varying error surface, a "frame of reference" for optimal adaptation.

The update steps (indicated by the index $k|k$) and the prediction steps (index $k+1|k$) for all the quantities involved are defined below as

$$\begin{aligned}
\tilde{\boldsymbol{w}}_{k|k} &\overset{\text{def}}{=} \boldsymbol{w}_k^{\text{o}} - \boldsymbol{w}_{k|k},\\
\mathbf{P}_{k|k} &\overset{\text{def}}{=} E\{\tilde{\boldsymbol{w}}_{k|k}\tilde{\boldsymbol{w}}_{k|k}^{\mathrm{T}}\},\\
\tilde{\boldsymbol{w}}_{k+1|k} &\overset{\text{def}}{=} \boldsymbol{w}_{k+1}^{\text{o}} - \boldsymbol{w}_{k+1|k} = \mathbf{F}_k\tilde{\boldsymbol{w}}_{k|k} + \mathbf{q}_k,\\
\mathbf{P}_{k+1|k} &\overset{\text{def}}{=} E\{\tilde{\boldsymbol{w}}_{k+1|k}\tilde{\boldsymbol{w}}_{k+1|k}^{\mathrm{T}}\}\\
&= \mathbf{F}_k\mathbf{P}_{k|k}\mathbf{F}_k^{\mathrm{T}} + \mathbf{Q}_s. \qquad (24)
\end{aligned}$$

Much like (13)–(17), the Kalman gain is derived based on the weight error vector recursion, obtained by subtracting the optimal time-varying $\boldsymbol{w}_k^{\text{o}}$ from the state update in (23a), to yield

$$\tilde{\boldsymbol{w}}_{k|k} = \tilde{\boldsymbol{w}}_{k|k-1} - \mathbf{g}_k\boldsymbol{x}_k^{\mathrm{T}}\tilde{\boldsymbol{w}}_{k|k-1} - \mathbf{g}_k n_k, \quad (25)$$

so that the evolution of the weight error covariance becomes

$$\begin{aligned}
\mathbf{P}_{k|k} &\overset{\text{def}}{=} E\{\tilde{\boldsymbol{w}}_{k|k}\tilde{\boldsymbol{w}}_{k|k}^{\mathrm{T}}\}\\
&= \mathbf{P}_{k|k-1} - (\mathbf{P}_{k|k-1}\boldsymbol{x}_k\mathbf{g}_k^{\mathrm{T}} + \mathbf{g}_k\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k|k-1})\\
&\quad + \mathbf{g}_k\mathbf{g}_k^{\mathrm{T}}(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k|k-1}\boldsymbol{x}_k + \sigma_n^2). \qquad (26)
\end{aligned}$$

Finally, the Kalman gain, $\mathbf{g}_k$, which minimizes the MSD, $J_{k|k} = \mathrm{tr}\{\mathbf{P}_{k|k}\}$, is obtained as [1]

$$\mathbf{g}_k = \frac{\mathbf{P}_{k|k-1}}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k|k-1}\boldsymbol{x}_k + \sigma_n^2}\boldsymbol{x}_k = \mathbf{G}_k\boldsymbol{x}_k. \quad (27)$$

which is conformal with the optimal LMS gain in (19). The general Kalman filter steps are summarized in Algorithm 3.

REMARK 5

Steps 1–3 in Algorithm 3 are identical to the deterministic Kalman filter that was derived starting from the LMS and is described in Algorithm 1. The essential difference is in steps 4 and 5, which cater for the time-varying and stochastic general system weights. Therefore, the fundamental principles of the Kalman filter can be considered through optimal adaptive step-size LMS algorithms.

Algorithm 3: The general Kalman filter.

```
At each time instant $k > 0$, based on measurements $\{d_k, \boldsymbol{x}_k\}$
  1) Compute the optimal learning gain (Kalman gain):
        $\mathbf{g}_k = \mathbf{P}_{k|k-1}\boldsymbol{x}_k/(\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k|k-1}\boldsymbol{x}_k + \sigma_n^2)$
  2) Update the weight vector estimate:
        $\boldsymbol{w}_{k|k} = \boldsymbol{w}_{k|k-1} + \mathbf{g}_k(d_k - \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{w}_{k|k-1})$
  3) Update the weight error covariance matrix:
        $\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{g}_k\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}$
  4) Predict the next (posterior) weight vector (state):
        $\boldsymbol{w}_{k+1|k} = \mathbf{F}_k\boldsymbol{w}_{k|k}$
  5) Predict the weight error covariance matrix:
        $\mathbf{P}_{k+1|k} = \mathbf{F}_k\mathbf{P}_{k|k}\mathbf{F}_k^{\mathrm{T}} + \mathbf{Q}_s$
```

[TABLE 2] A SUMMARY OF OPTIMAL GAIN VECTORS. THE OPTIMAL STEP-SIZES FOR THE LMS-TYPE ALGORITHMS ARE LINKED TO THE A PRIORI VARIANT OF THE KALMAN GAIN VECTOR, $\mathbf{g}_k$, SINCE $\mathbf{P}_{k|k-1} = \mathbf{P}_{k-1}$ FOR DETERMINISTIC AND TIME-INVARIANT SYSTEM WEIGHT VECTORS.

ALGORITHM	GAIN VECTOR	OPTIMAL GAIN VECTOR
KALMAN FILTER	$\mathbf{g}_k$	$\frac{\mathbf{P}_{k\|k-1}\boldsymbol{x}_k}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k\|k-1}\boldsymbol{x}_k + \sigma_n^2}$
LMS NLMS	$\mu_k\boldsymbol{x}_k$ $\rho_k\frac{\boldsymbol{x}_k}{\lVert\boldsymbol{x}_k\rVert^2}$	$\frac{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k}{\boldsymbol{x}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{x}_k + \sigma_n^2}\frac{\boldsymbol{x}_k}{\lVert\boldsymbol{x}_k\rVert^2}$
$\varepsilon -$ NLMS	$\frac{\boldsymbol{x}_k}{\lVert\boldsymbol{x}_k\rVert^2 + \varepsilon_k}$	which equals $\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{g}_k\frac{\boldsymbol{x}_k}{\lVert\boldsymbol{x}_k\rVert^2}$

CONCLUSIONS

We have employed "optimal gain" as a mathematical lens to examine conjointly variants of the LMS algorithms and Kalman filters. This perspective enabled us to create a framework for unification of these two main classes of adaptive recursive online estimators. A close examination of the relationship between the two standard performance evaluation measures, the MSE and MSD, allowed us to intuitively link up the geometry of learning of Kalman filters and LMS, within both deterministic and stochastic system identification settings. The Kalman filtering algorithm is then derived in an LMS-type fashion via the optimal learning gain matrix, without resorting to probabilistic approaches [12].

Such a conceptual insight permits seamless migration of ideas from the state-space-based Kalman filters to the LMS adaptive linear filters and vice versa and provides a platform for further developments, practical applications, and nonlinear extensions [13]. It is our hope that this framework of examination of these normally disparate areas will both demystify recursive estimation for educational purposes [14], [15] and further empower practitioners with enhanced intuition and freedom in algorithmic design for the manifold applications.

ACKNOWLEDGMENTS

We thank Prof. Jonathon Chambers, Prof. Ljubisa Stankovic, and Dr. Vladimir Lucic for their insightful comments. This material was finely tuned through our lecture courses on statistical signal processing at Imperial College, notwithstanding the pivotal feedback from our students.

AUTHORS

Danilo P. Mandic (d.mandic@imperial.ac.uk) is a professor of signal processing at Imperial College London, United Kingdom. He has been working in statistical signal processing and specializes in multivariate

state-space estimation and multidimensional adaptive filters. He received the President Award for excellence in postgraduate supervision at Imperial College in 2014. He is a Fellow of the IEEE.

Sithan Kanna (shri.kanagasabapathy08@imperial.ac.uk) is a Ph.D. candidate in statistical signal processing at Imperial College London, United Kingdom, and has been working in state-space estimation and adaptive filtering. He was awarded the Rector's Scholarship at Imperial College.

Anthony G. Constantinides (a.constantinides@imperial.ac.uk) is Emeritus Professor at Imperial College London, United Kingdom. He is a pioneer of signal processing and has been actively involved in research on various aspects of digital signal processing and digital communications for more than 50 years. He is a Fellow of the Royal Academy of Engineering and the 2012 recipient of the IEEE Leon K. Kirchmayer Graduate Teaching Award. He is a Fellow of the IEEE.

REFERENCES

[1] D. J. Simon, *Optimal State Estimation: Kalman, H-Infinity and Non-Linear Approaches*. Hoboken, NJ: Wiley, 2006.

[2] D. Williams, *Probability with Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[3] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1985.

[4] V. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. Signal Processing*, vol. 41, pp. 2075–2087, June 1993.

[5] S. Douglas, "Generalized gradient adaptive step sizes for stochastic gradient adaptive filters," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 1995, vol. 2, pp. 1396–1399.

[6] S. Douglas, "A family of normalized LMS algorithms," *IEEE Signal Processing Lett.*, vol. 1, no. 3, pp. 49–51, 1994.

[7] C. G. Lopes and J. Bermudez, "Evaluation and design of variable step size adaptive algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2001, vol. 6, pp. 3845–3848.

[8] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ: Prentice Hall, 2000.

[9] A. Sayed and T. Kailath, "A state-space approach to adaptive RLS filtering," *IEEE Signal Processing Mag.*, vol. 11, pp. 18–60, July 1994.

[10] D. Mandic, "A generalized normalized gradient descent algorithm," *IEEE Signal Processing Lett.*, vol. 11, pp. 115–118, Feb. 2004.

[11] J. Fernandez-Bes, V. Elvira, and S. Van Vaerenbergh, "A probabilistic least-mean-squares filter," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2199–2203.

[12] R. Faragher, "Understanding the basis of the Kalman filter via a simple and intuitive derivation," *IEEE Signal Processing Mag.*, vol. 29, no. 5, pp. 128–132, 2012.

[13] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures, and Stability*. New York: Wiley, 2001.

[14] J. Humpherys and J. West, "Kalman filtering with Newton's method," *IEEE Control Syst. Mag.*, vol. 30, no. 6, pp. 49–51, 2010.

[15] A. Nehorai and M. Morf, "A mapping result between Wiener theory and Kalman filtering for nonstationary," *IEEE Trans. Automat. Contr.*, vol. 30, no, 2, pp. 175–177, Feb. 1985.

[SP]

Digital Object Identifier 10.1109/MSP.2015.2484400