An Interpretation of Kalman Filter from Least Mean Square Algorithm

Zhiqiang Xie
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
xiezhq@shanghaitech.edu.cn

Abstract

This is the final report of the project on EE251 ***Signal Detection, Parameter Estimation, and Statistical Learning*** *course. Through this report, we'll interpret Kalman Filter from a least mean square (LMS) algorithm view, and show the intrinsic relationship between these two popular adaptive estimation algorithm. A generic system identification problem is introduced to demonstrate this idea. We'll see that the Kalman gain is precisely the optimal learning gain for LMS algorithm from specific problem to general case, and Kalman filter can be interpreted as a LMS algorithm with optimal step size.*

1. Introduction

The Kalman filter and the least mean square (LMS) adaptive filter are two of the most popular adaptive estimation algorithms which are widely used in many signal processing systems. They are typically treated as separate entities. The former one comes from Bayesian sequential estimator and the latter one is always associated with gradient descent method, which is a general training method for unknown parameters in model.

In other words, Kalman filter search targets from the state space but the gradient decent search in a space with higher dimension. Prior knowledge is the key element for Bayesian estimators including Kalman filter, that's the crucial difference between Kalman filter and LMS algorithm.

In this report, we'll not talk about Bayesian statistics, but a straightforward view of optimal parameter selection in learning algorithm. The derivation procedures are based on this note[5]. We'll finally show that the Kalman gain is precisely the optimal learning gain for LMS algorithms, and Kalman filter can be interpreted as a LMS algorithm with optimal step size.

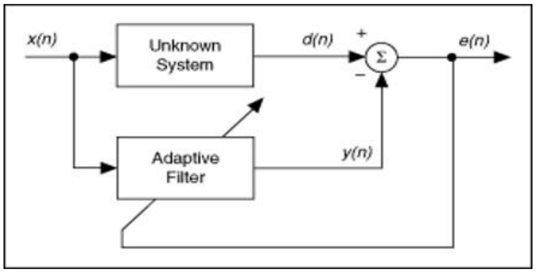

Figure 1. System identification

2. Problem Formulation

System identification is a methodology for building mathematical models of dynamic systems using measurements of the system's input and output signals. This methodology provides

Algorithm	d_k	x_k	w_k^o
Kalman Filter	Measurement	Observation Model	State
LMS	Desired Output	Input Data	True Weight

Table 1. Different Interpretation of Notations

a fundamental view for many problems like active noise control (ANC)[2]. A typical diagram of system identification is shown in figure 1. We introduce a generic system identification setting to talk about the Kalman filter and LMS algorithm:

$$d_k = x_k^T w_k^o + n_k \tag{1}$$

where the equation aims to estimate the unknown system parameters w_k^o.

2.1. Preliminaries

As for equation (1), here is the list of notations:

1. $\mathbf{d_k}$, observation of the system to be estimated
2. $\mathbf{x_k}$, input to the unknown system at time k
3. $\mathbf{w_k^o}$, target system weight at time k
4. $\mathbf{n_k}$, measurement noise at time k

where x_k is usually designated as 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. We'll first discuss the case with time-invariant target weight $w_k^o = w^o$, and then extend it to general case.

For clarification of different interpretation of notations, we summarize them in Table 1.

2.2. Formulation

From the system response equation:

$$\begin{aligned} d_k =& x_k^T w^o + n_k \\ \hat{w^o} =& w_k = f(w_{k-1}, d_k, x_k) \end{aligned}$$

where f is the estimator, thus we have the MSE criteria used in adaptive algorithm:

$$\begin{aligned} e_k =& d_k - x_k^T w_{k-1} \\ MSE =& M_k = E\{e_k^2\} \end{aligned}$$

Here we can find the innovation defined in Kalman filter and the covariance matrix of it:

$$\begin{aligned} \tilde{w}_k =& w^o - w_k \\ P_k =& E\{\tilde{w}_k \tilde{w}_k^T\} \\ e_k =& x_k^T \tilde{w}_{k-1} + n_k \end{aligned}$$

The innovation $\tilde{w}_k$ here is no doubt to be one of the best criteria for this system identification problem, because it directly indicates the difference between our estimation and the target. However, we already know that $E\{||\tilde{w}_k||^2\}$ is the Bayesian MSE, which we want to minimize in the iterations of Kalman filter, we'll call it MSD (mean square deviation).

Moreover, we shall introduce some more relations:

$$\begin{aligned} J_k =& E\{||\tilde{w}_k||^2\} \\ =& tr\{P_k\} \\ M_k =& E\{(x_k^T \tilde{w}_{k-1} + n_k)^2\} \\ =& x_k^T P_{k-1} x_k + \sigma_n^2 \end{aligned}$$

where we use J_k to represent the MSD criteria we adopt in Kalman filter, and the M_k is the MSE criteria for LMS algorithm. Besides, it should be clear that J_k is closely related to M_k: minimizing the MSD also corresponds to minimizing the MSE. And almost all adaptive methods for this system identification problem does the same thing like this:

$$\begin{aligned} w_k =& w_{k-1} + \Delta w_k \\ =& w_{k-1} + g_k e_k \end{aligned}$$

where g_k is the learning gain for every error.

3. Intrinsic Relationship

3.1. Optimal Scalar Step Size for LMS

The LMS algorithm often employs stochastic gradient decent (SGD) to approximately minimize the MSE M_k:

$$w_k = w_{k-1} + \mu_k x_k e_k$$

where the parameter μ_k is the step size which may be crucial for the convergence of adaptive methods. The parameter is supposed to be varying for different accuracy need, and some standard approaches find good step size that minimize the MSE via $\partial M_k / \partial \mu_k$[6].

Our idea is to introduce an optimal step size into LMS algorithm based on the direct minimization of the J_k. From the recursion relation, we have:

$$\begin{aligned}\tilde{w}_k =& \tilde{w}_{k-1} - g_k x_k^T \tilde{w}_{k-1} - g_k n_k \\ P_k =& E\{\tilde{w}_k \tilde{w}_k^T\} \\ =& P_{k-1} - (P_{k-1} x_k g_k^T + g_k x_k^T P_{k-1}) \\ &+ g_k g_k^T (x_k^T P_{k-1} x_k + \sigma_n^2)\end{aligned}$$

As we know:

$$\begin{aligned}tr\{P_{k-1} x_k g_k^T\} =& tr\{g_k x_k^T P_{k-1}\} \\ =& g_k^T P_{k-1} x_k\end{aligned}$$

a similar equation about J_k holds:

$$J_k = J_{k-1} - 2 g_k^T P_{k-1} x_k \quad (2)$$

$$+ ||g_k||^2 (x_k^T P_{k-1} x_k + \sigma_n^2) \quad (3)$$

By substituting the gain $g_k = \mu_k x_k$ into the equation (2):

$$\begin{aligned}J_k =& J_{k-1} - 2\mu_k x_k^T P_{k-1} x_k \\ &+ \mu_k^2 ||x_k||^2 (x_k^T P_{k-1} x_k + \sigma_n^2)\end{aligned}$$

As for the above equation, we can obtain the optimal step size which minimizes J_k by solving for μ_k via $\partial J_k / \partial \mu_k = 0$. It yields[4]:

$$\mu_k = \frac{1}{||x_k||^2} \frac{x_k^T P_{k-1} x_k}{x_k^T P_{k-1} x_k + \sigma_n^2}$$

here μ_k is the optimal step size for LMS.

Algorithm 1 The Kalman filter for deterministic states

```
At each time instant $k > 0$, based on measurements $\{d[k], x[k]\}$
1) Compute the Kalman gain (optimal learning gain):
    $g[n] = P_{k-1} x_k / (x_k^T P_{k-1} x_k + \sigma_n^2)$
2) Update the weight estimate:
    $w_k = w_{k-1} + g_k (d_k - x_k^T w_{k-1})$
3) Update the innovation covariance matrix (predict error):
    $P_k = P_{k-1} - g_k x_k^T P_{k-1}$
```

3.2. From LMS to Kalman Filter

As we mentioned, the criteria MSD is actually the Bayesian MSE for Kalman filter. Here we solve the equation (2) for g_k via $\partial J_k / \partial g_k = 0$, the optimal gain for LMS emerges:

$$g_k = \frac{P_{k-1}}{x_k^T P_{k-1} x_k + \sigma_n^2} x_k = G_k x_k$$

Here the optimal gain is precisely the Kalman gain[7], and the overall procedures of Kalman filter for this problem is described in Algorithm 1. You may notice that the predict phases for the state and MSE are missing, as for the prediction:

$$\hat{w}_{k|k-1} = w_{k-1}$$

which means the states tell nothing (if we know nothing about the varying, we can't do anything but treat it as deterministic) about next prediction. However, for deterministic states: the equation is correct and accurate.

Specially, all derivations above come from a LMS view instead of sequential Bayesian estimation. It's clear that Kalman filter can be interpreted as a LMS algorithm with optimal learning gain for scalar case yet. Furthermore, it's also

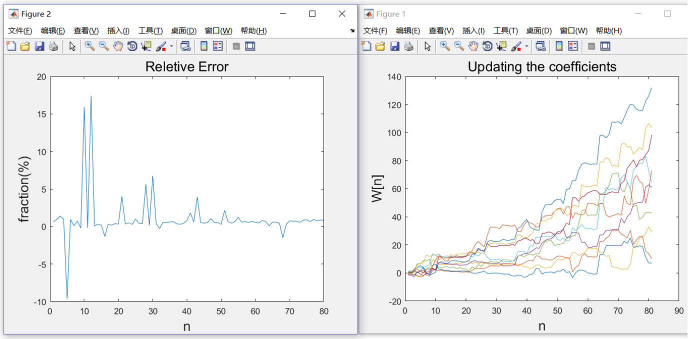

Figure 2. The Simulation Result of Time Varying Case

suited to variants of LMS, such as NLMS (normalized LMS) and ε-NLMS:

$$\begin{aligned} w_k =& w_{k-1} + \rho_k \frac{x_k}{||x_k||^2} e_k \\ w_k =& w_{k-1} + \frac{x_k}{||x_k||^2 + \varepsilon_k} e_k \end{aligned}$$

where $\rho = \frac{x_k^T P_{k-1} x_k}{x_k^T P_{k-1} x_k + \sigma_n^2}$ and $\varepsilon_k = \frac{||x||^2 \sigma_n^2}{x_k^T P_{k-1} x_k}$ can result in the equivalence.

3.3. From Optimal LMS to General Kalman Filter

If the system weight we want to estimate is time-varying, here comes the general case:

$$\begin{aligned} w_{k+1}^o =& F_k w_k^o + q_k \\ d_k =& x_k^T w_k^o + n_k \end{aligned}$$

where $q_k \sim \mathcal{N}(0, Q_s)$ and $n_k \sim \mathcal{N}(0, \sigma_n^2)$.

Now $w_{k|k-1} \neq w_{k-1}$, the new rules are:

$$\begin{aligned} w_{k|k} =& w_{k|k-1} + g_k(d_k - x_k^T w_{k|k-1}) \\ w_{w+1|k} =& F_k w_{k|k} \\ \tilde{w}_{k+1|k} =& F_k \tilde{w}_{k|k} + q_k \\ P_{k+1|k} =& E \tilde{w}_{k+1|k} \tilde{w}_{k+1|k}^T \\ =& F_k P_{k|k} F_k^T + Q_s \end{aligned}$$

Actually, here are less thing new for the general case but updating the prediction before whenever we do correction.

4. Comparison & Experiments

Since we have shown Kalman filter performing optimal step size, it converges faster than many

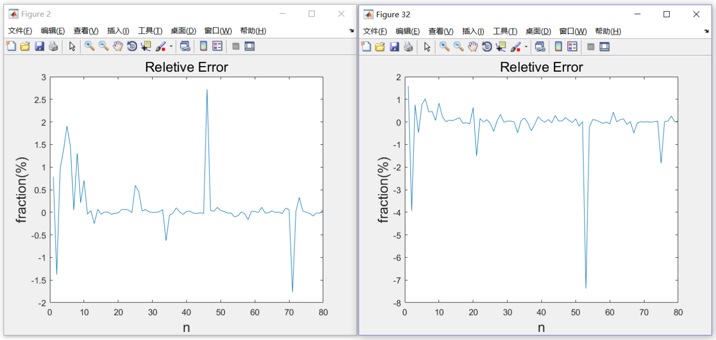

Figure 3. 3 times (left) and 6 times (right) Deviation

adaptive algorithms like LMS. Generally, we'll choose Kalman Filter if we have sufficient prior knowledge from system.

As for the time varying system identification problem, we usually don't know about the specific rate of varying. Actually, a Kalman filter without the information about varying can work well either owing to its fast convergence. The figure 2 depicted a system with 5% increasing rate, and we can see the MSE can be kept in a low level, either.

However, deficiency of prior knowledge is not worse than the deviation of the known prior knowledge. We introduce 3 times and 6 times deviation of prior knowledge σ_n^2 respectively, the simulation result is shown in figure 3. Note the difference of scale of y-axis, it's clear that Kalman filter is not reliable without good prior knowledge. In practice, good prior knowledge is not easy to obtain. That's why purely data-driven methods are more and more popular nowadays.

5. Conclusion and Related Work

By selecting a special criterion for optimization problem, we can derive the Kalman filtering algorithm in an LMS-type fashion via the optimal learning gain matrix, without resorting to probabilistic approaches[1].

Nowadays, with the development of DSP technique, the calculation speed (time per operation) is not the bottleneck of the performance of many signal processing application based on adaptive algorithm. More and more commercial products care more about adaptability (less prior knowledge and fast convergence), which encourage many data-driven methods with fast convergence were proposed. However, since Kalman filter has been proved to be the optimal (fast) deterministic algorithm, some researchers are trying to overcome its drawbacks.

For instance, a purely data-driven Kalman-like (no prior knowledge but estimating online) algorithm for ANC problem is proposed[3] recently, which achieved a convincing performance. That might be a bright direction of further ANC research.

It could be helpful to analyze the fundamental methods from different views, and then we might know the drawbacks as well the potential of these approaches, which may guide us a bright way to a better solution.

References

[1] R. Faragher. Understanding the basis of the kalman filter via a simple and intuitive derivation [lecture notes]. *IEEE Signal processing magazine*, 29(5):128–132, 2012.

[2] S. M. Kuo and D. R. Morgan. Active noise control: a tutorial review. *Proceedings of the IEEE*, 87(6):943–973, 1999.

[3] S. Liebich, J. Fabry, P. Jax, and P. Vary. Time-domain kalman filter for active noise cancellation headphones. In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 593–597. IEEE, 2017.

[4] C. G. Lopes and J. M. Bermudez. Evaluation and design of variable step size adaptive algorithms. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 6, pages 3845–3848. IEEE, 2001.

[5] D. P. Mandic, S. Kanna, and A. G. Constantinides. On the intrinsic relationship between the least mean square and kalman filters [lecture notes]. *IEEE Signal Processing Magazine*, 32(6):117–122, 2015.

[6] V. J. Mathews and Z. Xie. A stochastic gradient adaptive filter with gradient adaptive step size. *IEEE Transactions on Signal Processing*, 41(6):2075–2087, 1993.

[7] S. K. Sengijpta. Fundamentals of statistical signal processing: Estimation theory, 1995.