MANCHESTER 1824

The University of Manchester

Discussion Paper Series

A note on exact correspondences between adaptive learning algorithms and the Kalman filter

By

Michele Berardi and Jaqueson K. Galimberti

Centre for Growth and Business Cycle Research, Economic Studies,
University of Manchester, Manchester, M13 9PL, UK

June 2012
Number 170

Download paper from:

http://www.socialsciences.manchester.ac.uk/cgbcr/discussionpapers/index.html

A note on exact correspondences between adaptive learning algorithms and the Kalman filter

Michele Berardi

The University of Manchester

Jaqueson K. Galimberti*

The University of Manchester and The Capes Foundation

June 20, 2012

Abstract

Digressing into the origins of the two main algorithms considered in the literature of adaptive learning, namely Least Squares (LS) and Stochastic Gradient (SG), we found a connection between their non-recursive forms and their interpretation within a state-space unifying framework. Based on such connection, we extend the correspondence between the LS and the Kalman filter recursions to a formulation with time-varying gains of the former, and also present a similar correspondence for the case of the SG. Our correspondences hold exactly, in a computational implementation sense, and we discuss how they relate to previous approximate correspondences found in the literature.

Keywords: adaptive learning, least squares, stochastic gradient, Kalman filter.

JEL codes: C32, C63, D83, D84.

1 Introduction

Adaptive learning algorithms have been proposed to provide the heuristics through which agents can be assumed to form their expectations, in order to pragmatically validate whether the consistency requirements inherent to rational expectations (RE) can be satisfied by boundedly rational agents (see Evans and Honkapohja, 2001). Going beyond the RE paradigm, however, comes at the cost of introducing another degree of freedom into the analysis, as one (or more) learning algorithm(s) must be specified to represent the evolution of agents beliefs. Two algorithms have received most of the attention in the literature, namely, Least Squares (LS) and Stochastic Gradient (SG), and the different formulations from which these algorithms can be obtained as estimators is the subject of this note.

*Corresponding author. E-mail: jaqueson.galimberti@postgrad.manchester.ac.uk.

Since the seminal works in the subject of learning and expectations in macroeconomics (Bray, 1982; Marcet and Sargent, 1989) the LS algorithm has been taken as the natural choice to represent agents mechanism of adaptive learning. This choice is in general attributed to the widespread knowledge of its ordinary counterpart, the so-called Ordinary Least Squares (OLS) estimator, between econometricians. The SG algorithm, on the other hand, provides a computationally simpler alternative, leading some authors to advocate for its use as a more plausible learning device from a bounded rationality standpoint (Barucci and Landi, 1997; Evans and Honkapohja, 1998).

Although the LS and the SG algorithms share a similar recursive formulation, which makes them suitable for mimicking agents adjusting their forecasts as new data becomes available over time, their recursions can be clearly distinguished for the application, or not, respectively, of a "normalization" step during the updating process. It is also for the absence of this specific mechanism that the SG is characterized by a lower computational complexity as compared to the LS. Apart from this computational difference, for each of these algorithms one can also adopt distinctive formulations with respect to their learning gain, which is a parameter determining how quickly a given information is incorporated into the algorithm's coefficients estimates. Three of the main alternatives for the specification of this learning gain are those of a time-decreasing, a time-constant, and a time-varying (not restricted to be decreasing) sequence of values, and their suitability depends on the time-varying nature of the environment.

A decreasing-gain LS was the seminal choice in the learning literature, so as to match the recursive form of the OLS estimator. For the case of linear models with time-invariant parameters, this estimator is known to possess some well desired properties, such as consistency and efficiency, though these properties do not extend to a time-varying context. This latter fact implies the intriguing observation that a decreasing-gain LS learning meachanism is appropriate only along the time-invariant path of a RE equilibrium, where learning itself is indeed pointless (Bray and Savin, 1986).

Extensive evidence (see Stock and Watson, 2003; Cogley and Sargent, 2005; Sims and Zha, 2006; Sargent et al., 2006) favoring time-varying parameter models of the economy has, nevertheless, challenged this paradigm, and the departure from the parameter constancy assumption (see Margaritis, 1990; Bullard, 1992; McGough, 2003) has naturally led to the requirement of adjustments to the learning rules as well. These adjustments came first into the form of constant-gain learning (Sargent, 1999), and later into the more general form of time-varying sequences of learning gains (Marcet and Nicolini, 2003).

One way to deepen our understanding of these learning rules has been through the establishment of correspondences between them and the Kalman filter (as done in Ljung and Soderstrom, 1983; Ljung and Gunnarsson, 1990; Sargent, 1999; Sargent and Williams, 2005; Branch and Evans, 2006; Evans et al., 2010), for which optimality properties are known from a long standing literature (see Anderson and Moore, 1979). Previous studies, however, have focused mainly on the analysis of the LS algorithm (Sargent, 1999; Branch and Evans, 2006), while correspondence results for the SG case have been found

to hold only approximately in a long-run sense, where any transient phase affecting the algorithm's estimates has already died out (Sargent and Williams, 2005; Evans et al., 2010). Furthermore, these correspondences have been separately drawn for the specific cases of predefined decreasing and constant sequences of gains.

It is the purpose of this note to extend the exact correspondence results for the LS algorithm both to the case of the SG algorithm, as well as to the more general case of an unrestricted time-varying sequence of learning gains. We do that by providing a renewed interpretation of how these correspondences can be drawn with respect to the non-recursive forms of the LS and the SG algorithms, presented in Section 2, where the non-recursive form for the latter algorithm is also an original feature of this note. As we adopt an exact approach in drawing our correspondences, instead of the above mentioned approximated sense, we argue that our results favor both the computational implementation of these algorithms, as well as their employment over out-of-equilibrium paths. We present our exact correspondences in Section 3, while a discussion about how they relate to previous approximate correspondences is postponed to Section 4. Section 5 then concludes.

2 Digression into algorithms origins

2.1 Preliminaries

To understand the differences between the LS and the SG origins lets first establish a common context of estimation. Our focus here is on linear regressions of the form[1]

$$y_t = \mathbf{x}_t'\boldsymbol{\theta}_t + \varepsilon_t, \tag{1}$$

where y_t is assumed to be related to a vector of (pre-determined) variables $\mathbf{x}_t = (x_{1,t}, \ldots, x_{K,t})'$ through the vector of (possibly time-varying) coefficients $\boldsymbol{\theta}_t = (\theta_{1,t}, \ldots, \theta_{K,t})'$, and ε_t denotes a (Gaussian) white noise disturbance with variance given by $\sigma_t^2 = E\left[\varepsilon_t^2\right]$. Interest is on the estimation of $\boldsymbol{\theta}_t$ with given observations of y_t and $\mathbf{x}_t$, but not of ε_t.

Under this context, the LS algorithm we are interested in assumes the form of

$$\hat{\boldsymbol{\theta}}_t^{LS} = \hat{\boldsymbol{\theta}}_{t-1}^{LS} + \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\theta}}_{t-1}^{LS}\right), \tag{2}$$

$$\mathbf{R}_t = \mathbf{R}_{t-1} + \gamma_t\left(\mathbf{x}_t\mathbf{x}_t' - \mathbf{R}_{t-1}\right), \tag{3}$$

where γ_t is a learning gain parameter, and $\mathbf{R}_t$ stands for an estimate of regressors' matrix of second

[1] Our results can be straightforwardly extended to a multivariate regressions context, an autoregressive context, or yet in both dimensions to a VAR specification.

moments, $E[\mathbf{x}_t\mathbf{x}_t']$. Under the same context, the SG algorithm is given by

$$\hat{\boldsymbol{\theta}}_t^{SG} = \hat{\boldsymbol{\theta}}_{t-1}^{SG} + \mu_t \mathbf{x}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{t-1}^{SG} \right), \tag{4}$$

with μ_t standing for the learning gain parameter in this case. The hat in $\hat{\boldsymbol{\theta}}_t^{LS}$ and $\hat{\boldsymbol{\theta}}_t^{SG}$ indicates that they stand for estimates of $\boldsymbol{\theta}_t$ in (1), based on period t information.

The main difference between the LS and the SG origins resides in how the estimation problem was first formulated, either into a non-recursive (block) minimization problem or into a recursive (filtering) form, respectively. In spite of this distinction, each algorithm can be interpreted under both formulations.

2.2 Non-recursive forms

The LS is originally derived from a non-recursive estimation problem (see Ljung and Soderstrom, 1983, pp. 57-61), namely the minimization of the sum of weighted error squares as given by

$$\hat{\boldsymbol{\theta}}_t^{LS} = \arg\min \sum_{i=1}^{t} \beta(t,i) \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\theta}}_t^{LS} \right)^2, \tag{5}$$

where

$$\beta(t,i) = \begin{cases} \alpha_i \prod_{k=i+1}^{t} \lambda_k & for\ i < t, \\ \alpha_i & for\ i = t, \end{cases} \tag{6}$$

indicates how past observations are discounted, and thus, it is typically increasing in i for a given t. The structure in the weighting scheme imposed by (6), though not required, provides an exponential forgetting profile in the criterion (5) when $\lambda_k \leq 1$. In this sense, λ_k stands for the stream of forgetting factors and α_i regulates the ceiling to the weights these factors attach to observations. The solution to (5), which is quadratic in $\hat{\boldsymbol{\theta}}_t^{LS}$, results in the non-recursive LS estimator,

$$\hat{\boldsymbol{\theta}}_t^{LS} = \left[\sum_{i=1}^{t} \beta(t,i)\, \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i=1}^{t} \beta(t,i)\, \mathbf{x}_i y_i. \tag{7}$$

Letting

$$\lambda_k = \frac{\gamma_{k-1}}{\gamma_k}(1-\gamma_k) \ \ and\ \alpha_i = 1, \tag{8}$$

we obtain

$$\beta(t,i) = \begin{cases} \frac{\gamma_i}{\gamma_t} \prod_{k=i+1}^{t} (1-\gamma_k) & for\ i < t, \\ 1 & for\ i = t, \end{cases} \tag{9}$$

which indicates how the gains sequence in (2)-(3) gets translated into the weights put into past observations. It is particularly interesting to note that: (i) when $\gamma_i = 1/i$, $\beta(t,i) = 1$ and (7) reduces to the

OLS estimator; and, (ii) with a constant gain, $\gamma_i = \bar{\gamma}$, past observations receive geometrically decaying weights, i.e., $\beta(t,i) = (1-\bar{\gamma})^{t-i}$. Furthermore, it requires just a few derivations to show that the recursive form in (2)-(3) corresponds to the non-recursive solution in (7) with the weights given by (9)[2].

The SG algorithm can also be put into a similar non-recursive form. Using the notation of (6), the non-recursive SG is given by[3]

$$\hat{\boldsymbol{\theta}}_t^{SG} = \sum_{i=1}^{t} \boldsymbol{\beta}(t,i)\, \mathbf{x}_i y_i, \tag{10}$$

$$\boldsymbol{\lambda}_k = (\mathbf{I} - \mu_k \mathbf{x}_k \mathbf{x}_k') \ \textit{and}\ \alpha_i = \mu_i, \tag{11}$$

where we can see that compared to the LS non-recursive form, in (7), the SG does not have the "normalization" term given by the inverse of the regressors (sample) matrix of moments. On the other hand, the way the SG discounts past observations is not fully determined by the choice of the gains sequence as it happens for the LS, but it also depends on the data. This is clearly reflected into the definition of the forgetting factor in (11), which under a multivariate context turns itself, and the weighting factor in (10), into a matrix form. We take this latter difference as an explanation for the finding that the SG estimates are sensitive to data scales (Evans et al., 2010).

2.3 Filtering origins

The SG, in turn, is recursive from its origins (see Widrow, 1971): it stands as a stochastic approximation to the iterative method of Steepest Descent (SD) used to achieve the optimal solution to a linear filtering problem. Within the context of model (1) the filtering problem is to find a vector of coefficients $\hat{\boldsymbol{\theta}}_t$ such that the variance of the squared estimation error associated to these estimates is minimal, i.e.,

$$\hat{\boldsymbol{\theta}}_t = \arg\min \frac{1}{2} E\left[y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_t\right]^2, \tag{12}$$

the gradient of which leads to the first order condition

$$-E\left[\mathbf{x}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_t\right)\right] = \mathbf{0}, \tag{13}$$

which is famously known as the orthogonality condition, given that it states that optimality of $\hat{\boldsymbol{\theta}}_t$ requires that the estimation errors must be orthogonal to each regressor variable. Solving (13) for the coefficients vector we get to the optimal solution to the linear filtering problem,

$$E\left[\mathbf{x}_t \mathbf{x}_t'\right] \hat{\boldsymbol{\theta}}_t = E\left[\mathbf{x}_t y_t\right], \tag{14}$$

$$\hat{\boldsymbol{\theta}}_t = E\left[\mathbf{x}_t \mathbf{x}_t'\right]^{-1} E\left[\mathbf{x}_t y_t\right], \tag{15}$$

[2]See A.1.
[3]See A.2.

also known as the Wiener-Hopf equation. Notice that from a deterministic viewpoint, the use of averaged sample counterparts of the expectational operators in (15) would lead to a least squares solution resembling to (7).

From the stochastic viewpoint, if the covariance matrix of the regressors ($E[\mathbf{x}_t\mathbf{x}_t']$) and the cross-covariances between the regressors and the endogenous variable ($E[\mathbf{x}_t y_t]$) are known, the optimal solution can be readily computed from (15). Such a task, however, may become computationally cumbersome as the number of regressors increases. Furthermore, under the time-varying context of (1), these (cross-)covariances would be time-varying as well, and thus a new computation of (15) would be required for each new observation. A simpler alternative is to use a numerical optimization method in order to iteratively navigate along the error-performance surface, which is given by the objective function in the minimization of (12), until the optimal vector of coefficients is found. One such a method is the SD, which is developed to apply successive corrections to the coefficients estimates in the direction opposite to the gradient vector, (13), i.e.,

$$\hat{\boldsymbol{\theta}}_i^{SD} = \hat{\boldsymbol{\theta}}_{i-1}^{SD} + \kappa_i \left(E[\mathbf{x}_t y_t] - E[\mathbf{x}_t\mathbf{x}_t']\,\hat{\boldsymbol{\theta}}_{i-1}^{SD} \right), \tag{16}$$

where its is important to note that the coefficients estimates are not (necessarily) indexed by time, in the sense that the recursion can be applied more than once within the same set of information, and the parameter κ_i controls the size of the correction from one iteration to the next.

The similarity between the recursions adopted by the method of SD and the SG are not just a coincidence. As a matter of fact, the SG algorithm of (4) comes from a stochastic approximation rationale to the SD for the case where the relevant (cross-)covariances are not known. The idea is simply to replace the (theoretical) gradient in (16) by an estimate computed from the latest squared estimation error, i.e., the gradient of $\frac{1}{2}\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\theta}}_{i-1}\right)^2$ with respect to $\hat{\boldsymbol{\theta}}_{i-1}$. Doing that, one readily obtains the SG recursion of (4) from the stochastic approximation based on the SD of (16), noticing that the subscript indexes of the latter are converted to period (t) subscripts, and the step size parameter is converted to the learning gain parameter in the SG formulation.

A stochastic interpretation can also be given to the LS algorithm under the same filtering context that gives rise to the SG algorithm (Ljung and Soderstrom, 1983, pp. 46-8). Such a rationale is obtained by employing a method of iterative solution to (15) more sophisticated than the SD, namely the Newton's method. Although the same recursive structure of (16) is maintained, the difference is that under Newton's method the gradient is computed up to a second order expansion by multiplying the gradient derivative by the inverse of its associated Hessian matrix. Under (12) the Newton's method would be translated as

$$\hat{\boldsymbol{\theta}}_i^{Nw} = \hat{\boldsymbol{\theta}}_{i-1}^{Nw} + \eta_i E[\mathbf{x}_t\mathbf{x}_t']^{-1} \left(E[\mathbf{x}_t y_t] - E[\mathbf{x}_t\mathbf{x}_t']\,\hat{\boldsymbol{\theta}}_{i-1}^{Nw} \right). \tag{17}$$

Again, a recursive estimate for the Hessian matrix $E[\mathbf{x}_t\mathbf{x}_t']$ can be constructed by noting that under the target optimal solution, the Hessian matrix is the solution R to $E[\mathbf{x}_t\mathbf{x}_t' - R] = 0$. Applying the above iterative method to solve for this condition one obtains

$$\mathbf{R}_i = \mathbf{R}_{i-1} + \eta_i \left(E[\mathbf{x}_t\mathbf{x}_t'] - \mathbf{R}_{i-1} \right), \tag{18}$$

which after substitution of $E[\mathbf{x}_t\mathbf{x}_t']$ by its observed counterpart, $\mathbf{x}_t\mathbf{x}_t'$, and adjusting the subscripts and the gain, leads to the same recursion for the estimate of the matrix of second moments in the LS algorithm, (3). Substituting this estimate for the Hessian matrix in (17), and proceeding with the same approximations for the first order gradient as we did for the SD, one also finds that the stochastic version to the Newton's method in (17) resembles exactly the coefficients recursion of the LS in (2).

3 State-space unifying framework

3.1 State-space representation

Having established how the LS and the SG algorithms compare in terms of their original formulations, we now show how they both can be obtained as special cases of the Kalman filter when applied to the estimation of the time-varying parameters of the linear relationship assumed in (1). For that purpose we further assume these parameters follow a random walk model as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \tag{19}$$

where $\boldsymbol{\omega}_t$ is assumed to be (Gaussian) white noise with variances (and covariances) given by $\boldsymbol{\Omega}_t = E[\boldsymbol{\omega}_t\boldsymbol{\omega}_t']$. The random sequences ε_t and $\boldsymbol{\omega}_t$ are also assumed to be mutually independent. Notice that the Gaussianity assumptions on the distribution of these disturbances are required for the mean squares optimality of the Kalman filter.

Equations (1) and (19) are recognizably in a state-space form for a regression with time-varying coefficients, where the former is treated as the observation equation and the latter as the state equation (see Hamilton, 1994, pp. 372-408). The main advantage of such a state-space form is that it serves as a framework for the derivation of the Kalman filter used to obtain recursive estimates of the states based on the observed signals.

3.2 Kalman filter

Adapted to our context, the Kalman filtering recursion is given by[4]

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} + \mathbf{K}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{t-1} \right), \tag{20}$$

$$\mathbf{K}_t = \frac{\mathbf{P}_{t-1}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{t-1}\mathbf{x}_t + \sigma_t^2}, \tag{21}$$

$$\mathbf{P}_t = \left(\mathbf{I} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{t-1}\mathbf{x}_t + \sigma_t^2} \right) \mathbf{P}_{t-1} + \boldsymbol{\Omega}_t, \tag{22}$$

where $\mathbf{K}_t$ is known as the Kalman gain vector and $\mathbf{P}_t$ stands for the covariance matrix of the coefficients estimates, i.e., $\mathbf{P}_t = E\left[\left(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t \right) \left(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t \right)' \right]$.

Other than for its optimality properties, the Kalman filter also turns out to be useful for providing an unifying framework for the adaptive learning algorithms we are interested in. This is done by imposing restrictions on the dynamics of the second moments of the disturbances affecting the motion of the assumed state-space model, i.e., σ_t^2 and $\boldsymbol{\Omega}_t$ (see Ljung and Soderstrom, 1983; Ljung and Gunnarsson, 1990).

Our contribution here is to show that the specifications of σ_t^2 and $\boldsymbol{\Omega}_t$ that establish the LS and the SG algorithms as special cases of the Kalman filter can be connected to the definitions of the terms λ_t and α_t that we used in (8) to draw the correspondence between the recursive and the non-recursive forms of these algorithms. Namely, the general basis of the correspondences we are drawing here starts by assigning

$$\sigma_t^2 = \lambda_t \alpha_t^{-1}, \tag{23}$$

and then proceeding with a derivation of $\boldsymbol{\Omega}_t$ that would turn the Kalman filter recursions, (20)-(22), into the specific cases of the LS and of the SG algorithms, (2)-(3) and (4), respectively.

3.3 Least Squares

The LS algorithm can be obtained as the special case of the Kalman filter by setting

$$\sigma_t^2 = \frac{\gamma_{t-1}}{\gamma_t} \left(1 - \gamma_t \right), \tag{24}$$

$$\boldsymbol{\Omega}_t = \left(\frac{\gamma_t}{\gamma_{t-1}\left(1 - \gamma_t\right)} - 1 \right) \left(\mathbf{I} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{t-1}\mathbf{x}_t + \sigma_t^2} \right) \mathbf{P}_{t-1}. \tag{25}$$

[4] See A.3.

Substituting these quantities into (21) and (22), and adding a superscript to distinguish the resulting algorithm from the general Kalman filter, it is straightforward to find

$$\mathbf{K}_t^{LS} = \frac{\mathbf{P}_{t-1}^{LS}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{t-1}^{LS}\mathbf{x}_t + \frac{\gamma_{t-1}}{\gamma_t}(1-\gamma_t)}, \tag{26}$$

$$\mathbf{P}_t^{LS} = \frac{\gamma_t}{\gamma_{t-1}(1-\gamma_t)}\left(\mathbf{I} - \mathbf{K}_t^{LS}\mathbf{x}_t'\right)\mathbf{P}_{t-1}^{LS}, \tag{27}$$

which has the same form as the LS algorithm in (2)-(3) with the inversion of the matrix of second moments replaced by $\gamma_t^{-1}\mathbf{P}_t^{LS}$ using the matrix inversion lemma[5].

The above correspondence generalizes those of Ljung and Gunnarsson (1990, p. 10) and Sargent (1999, pp. 115-8) to the case of a time-varying gain.

3.4 Stochastic Gradient

The SG algorithm can be found as the special case of the Kalman filter when we set

$$\sigma_t^2 = \mu_t^{-1} - \mathbf{x}_t'\mathbf{x}_t, \tag{28}$$

$$\mathbf{\Omega}_t = \mathbf{I} - \left(\mathbf{I} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{t-1}\mathbf{x}_t + \sigma_t^2}\right)\mathbf{P}_{t-1}. \tag{29}$$

Substituting these into (21) and (22) we find that[6]

$$\mathbf{K}_t^{SG} = \mu_t\mathbf{x}_t, \tag{30}$$

$$\mathbf{P}_t^{SG} = \mathbf{I}. \tag{31}$$

Although this correspondence has been mentioned in Ljung and Gunnarsson (1990, p. 11), to the authors' knowledge its derivation from specific expressions for σ_t^2 and $\mathbf{\Omega}_t$ has never been made explicit in a reasonable sense into the previous literature. The only explicit derivation we have found so far was given by Karjalainen (1996, p. 34), which obtained this correspondence for the constant-gain SG. Although Karjalainen's derivation can be extended to the time-varying gain case, it suffers with two important drawbacks: (i) it requires a specific initialization of $\mathbf{P}_0$; and (ii) the computation of $\mathbf{P}_t^{SG}$, though not required for the computation of the SG coefficient estimates, is dependent on $t+1$ regressors' information, which is at odds with the main idea of recursive estimation.

[5] See A.4.
[6] See A.5.

4 Discussion

Our approach follows that of Ljung and Soderstrom (1983); Ljung and Gunnarsson (1990); Sargent (1999) where specific parametrizations of σ_t^2 and $\mathbf{\Omega}_t$ are hand-picked in order to make the algorithms match exactly the more general Kalman estimator applied to a state-space unifying framework that has been extensively explored in the empirical macroeconometrics literature (see Stock and Watson, 1996, and references therein). Following Benveniste et al. (1990), similar correspondences to the ones we obtain here have been drawn by Sargent and Williams (2005) and Evans et al. (2010). However, instead of holding exactly, their correspondences hold only in an approximated sense, when the algorithms transient phases have already died out.

From an applied standpoint, the main drawback of the approximate approach is that the accuracy of this approximation depends on how closer the initialization of the algorithm is to its steady state estimates. From the learning and expectations standpoint, the fact that the approximation holds only asymptotically makes it hard to use the resulting framework for a unifying analysis of both learning convergence and out-of-equilibrium dynamics. Our choice for exact correspondences, therefore, favors both the empirical applicability of the adaptive learning algorithms as well as their interpretation as learning devices operating, not necessarily but often, off the long run steady state path of inferences.

An understanding of the interplay between these features is taken as the main issue of Sargent and Williams (2005), which uses the above approximate framework to formalize the idea of agents learning priors about drifting coefficients. The asymptotic complementarity between convergence and escapes allowed these authors to isolate the influence of the priors over the occurrence of those distinct dynamical features, though these priors were taken as pre-determined. By generalizing the correspondences of the LS and SG learning algorithms with the Kalman filter, and further allowing for unrestricted time-varying gains, our results can be taken as providing a framework of analysis for a case under which agents are allowed to adapt their priors in accordance to their experience.

5 Conclusions

In this note we provided a renewed view on how the LS and the SG algorithms can be connected to the Kalman filter estimator under a context of regressions with time-varying parameters. Our approach innovates for being based on the use of similar non-recursive forms for these adaptive learning algorithms, from which we were able to derive their correspondences to the Kalman recursions under the general case of unrestricted time-varying learning gains. One special feature of our correspondences is that they hold exactly, instead of approximately in a long-run sense, and we argue that such feature favors both the computational implementation of these algorithms, as well as their interpretation as learning mechanisms operating off equilibrium paths.

References

Anderson, B.D.O., Moore, J.B., 1979. Optimal Filtering. Prentice-Hall, Englewood Cliffs, NJ.

Barucci, E., Landi, L., 1997. Least mean squares learning in self-referential linear stochastic models. Economics Letters 57, 313–317.

Benveniste, A., Metivier, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations. Srpinger-Verlag.

Branch, W.A., Evans, G.W., 2006. A simple recursive forecasting model. Economics Letters 91, 158–166.

Bray, M., 1982. Learning, estimation, and the stability of rational expectations. Journal of Economic Theory 26, 318–339.

Bray, M.M., Savin, N.E., 1986. Rational expectations equilibria, learning, and model specification. Econometrica 54, 1129–1160.

Bullard, J., 1992. Time-varying parameters and nonconvergence to rational expectations under least squares learning. Economics Letters 40, 159 – 166.

Cogley, T., Sargent, T.J., 2005. Drifts and volatilities: monetary policies and outcomes in the post wwii us. Review of Economic Dynamics 8, 262 – 302.

Evans, G.W., Honkapohja, S., 1998. Stochastic gradient learning in the cobweb model. Economics Letters 61, 333–337.

Evans, G.W., Honkapohja, S., 2001. Learning and expectations in macroeconomics. Frontiers of Economic Research, Princeton University Press, Princeton, NJ.

Evans, G.W., Honkapohja, S., Williams, N., 2010. Generalized stochastic gradient learning. International Economic Review 51, 237–262.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press.

Karjalainen, P.A., 1996. Estimation theoretical background of root tracking algorithms with applications to EEG. Ph.lic. thesis. University of Kuopio.

Ljung, L., Gunnarsson, S., 1990. Adaptation and tracking in system identification - a survey. Automatica 26, 7 – 21.

Ljung, L., Soderstrom, T., 1983. Theory and Practice of Recursive Identification. The MIT Press.

Marcet, A., Nicolini, J.P., 2003. Recurrent hyperinflations and learning. American Economic Review 93, 1476–1498.

Marcet, A., Sargent, T.J., 1989. Convergence of least squares learning mechanisms in self-referential linear stochastic models. Journal of Economic Theory 48, 337–368.

Margaritis, D., 1990. A time-varying model of rational learning. Economics Letters 33, 309 – 314.

McGough, B., 2003. Statistical learning with time-varying parameters. Macroeconomic Dynamics 7, 119–139.

Sargent, T., Williams, N., Zha, T., 2006. Shocks and government beliefs: The rise and fall of american inflation. American Economic Review 96, 1193–1224.

Sargent, T.J., 1999. The Conquest of American Inflation. Princeton University Press, Princeton, NJ.

Sargent, T.J., Williams, N., 2005. Impacts of priors on convergence and escapes from nash inflation. Review of Economic Dynamics 8, 360 – 391.

Sims, C.A., Zha, T., 2006. Were there regime switches in u.s. monetary policy? American Economic Review 96, 54–81.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14, 11–30.

Stock, J.H., Watson, M.W., 2003. Has the business cycle changed and why?, in: NBER Macroeconomics Annual 2002, Volume 17. National Bureau of Economic Research, Inc. NBER Chapters, pp. 159–230.

Widrow, B., 1971. Adaptive filters, in: Kalman, R., DeClaris, N. (Eds.), Aspects of Network and System Theory. Holt and Rinehart and Winston, Inc., New York, pp. 563–587.

A Detailed derivations of correspondences

A.1 Correspondence between non-recursive and recursive LS

First let

$$\mathbf{R}_t = \gamma_t \sum_{i=1}^{t} \beta(t,i)\, \mathbf{x}_i \mathbf{x}_i'. \tag{32}$$

from which we find that

$$\begin{aligned}
\mathbf{R}_t &= \gamma_t \sum_{i=1}^{t-1} \beta(t,i)\, \mathbf{x}_i \mathbf{x}_i' + \gamma_t \mathbf{x}_t \mathbf{x}_t', && (33)\\
&= (1-\gamma_t)\, \gamma_{t-1} \sum_{i=1}^{t-1} \beta(t-1,i)\, \mathbf{x}_i \mathbf{x}_i' + \gamma_t \mathbf{x}_t \mathbf{x}_t', && (34)\\
&= (1-\gamma_t)\, \mathbf{R}_{t-1} + \gamma_t \mathbf{x}_t \mathbf{x}_t', && (35)\\
\mathbf{R}_t &= \mathbf{R}_{t-1} + \gamma_t \left(\mathbf{x}_t \mathbf{x}_t' - \mathbf{R}_{t-1}\right), && (36)
\end{aligned}$$

where in the second step we used

$$\beta(t,i) = \frac{\gamma_{t-1}}{\gamma_t}(1-\gamma_t)\beta(t-1,i), \tag{37}$$

which comes directly from (9).

For the vector of coefficients, note that using (32) in (7) we have that

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_t^{LS} &= \gamma_t \mathbf{R}_t^{-1} \sum_{i=1}^{t} \beta(t,i)\mathbf{x}_i y_i, & (38)\\
&= \gamma_t \mathbf{R}_t^{-1} \left[\mathbf{x}_t y_t + \sum_{i=1}^{t-1} \beta(t,i)\mathbf{x}_i y_i \right], & (39)\\
&= \gamma_t \mathbf{R}_t^{-1} \left[\mathbf{x}_t y_t + \frac{\gamma_{t-1}}{\gamma_t}(1-\gamma_t) \sum_{i=1}^{t-1} \beta(t-1,i)\mathbf{x}_i y_i \right], & (40)\\
\hat{\boldsymbol{\theta}}_t^{LS} &= \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t y_t + (1-\gamma_t)\gamma_{t-1}\mathbf{R}_t^{-1} \sum_{i=1}^{t-1} \beta(t-1,i)\mathbf{x}_i y_i, & (41)
\end{aligned}$$

where in the third step we have again made use of (37). Now, note that lagging (38) one period and pre-multiplying it by $\mathbf{R}_t^{-1}\mathbf{R}_{t-1}$ we have that

$$\mathbf{R}_t^{-1}\mathbf{R}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^{LS} = \gamma_{t-1}\mathbf{R}_t^{-1} \sum_{i=1}^{t-1} \beta(t-1,i)\mathbf{x}_i y_i,$$

which can then be substituted into (41) leading us to

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_t^{LS} &= \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t y_t + (1-\gamma_t)\mathbf{R}_t^{-1}\mathbf{R}_{t-1}\hat{\boldsymbol{\theta}}_{t-1}^{LS}, & (42)\\
&= \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t y_t + (1-\gamma_t)\mathbf{R}_t^{-1}\left[\mathbf{R}_t(1-\gamma_t)^{-1} - \gamma_t(1-\gamma_t)^{-1}\mathbf{x}_t\mathbf{x}_t'\right]\hat{\boldsymbol{\theta}}_{t-1}^{LS}, & (43)\\
&= \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t y_t + \hat{\boldsymbol{\theta}}_{t-1}^{LS} - \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t\mathbf{x}_t'\hat{\boldsymbol{\theta}}_{t-1}^{LS}, & (44)\\
\hat{\boldsymbol{\theta}}_t^{LS} &= \hat{\boldsymbol{\theta}}_{t-1}^{LS} + \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\theta}}_{t-1}^{LS}\right), & (45)
\end{aligned}$$

where in the second step we made use of the expression preceding (36) above. Note that (36) and (45) assume exactly the same form as of (3) and (2), respectively, thus establishing the correspondence at scrutiny.

A.2 Correspondence between non-recursive and recursive SG

Lets first take out the last term of the summation in (10) resulting in

$$\hat{\boldsymbol{\theta}}_t^{SG} = \mu_t \mathbf{x}_t y_t + \sum_{i=1}^{t-1} \beta(t,i)\mathbf{x}_i y_i. \tag{46}$$

Now, notice that from (6) and (11) we have that

$$\begin{aligned}
\beta(t,i) &= \mu_i \prod_{k=i+1}^{t} \left(\mathbf{I} - \mu_k \mathbf{x}_k \mathbf{x}_k'\right), && (47)\\
&= \left(\mathbf{I} - \mu_t \mathbf{x}_t \mathbf{x}_t'\right) \mu_i \prod_{k=i+1}^{t-1} \left(\mathbf{I} - \mu_k \mathbf{x}_k \mathbf{x}_k'\right), && (48)\\
\beta(t,i) &= \left(\mathbf{I} - \mu_t \mathbf{x}_t \mathbf{x}_t'\right) \beta(t-1,i). && (49)
\end{aligned}$$

Substituting (49) into (46) we then have

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_t^{SG} &= \mu_t \mathbf{x}_t y_t + \left(\mathbf{I} - \mu_t \mathbf{x}_t \mathbf{x}_t'\right) \sum_{i=1}^{t-1} \beta(t-1,i)\, \mathbf{x}_i y_i, && (50)\\
&= \mu_t \mathbf{x}_t y_t + \left(\mathbf{I} - \mu_t \mathbf{x}_t \mathbf{x}_t'\right) \hat{\boldsymbol{\theta}}_{t-1}^{SG}, && (51)\\
\hat{\boldsymbol{\theta}}_t^{SG} &= \hat{\boldsymbol{\theta}}_{t-1}^{SG} + \mu_t \mathbf{x}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{t-1}^{SG}\right), && (52)
\end{aligned}$$

which has exactly the same form as of (4), thus establishing the correspondence at scrutiny.

A.3 Correspondence with general Kalman filter

Following Hamilton (1994, pp. 399-400) notation, we consider a general state-space model with stochastically varying coefficients given by

$$\begin{aligned}
\boldsymbol{\xi}_{t+1} &= \mathbf{F}(\mathbf{x}_t)\, \boldsymbol{\xi}_t + \mathbf{v}_{t+1}, && (53)\\
\mathbf{y}_t &= \mathbf{a}(\mathbf{x}_t) + [\mathbf{H}(\mathbf{x}_t)]'\, \boldsymbol{\xi}_t + \mathbf{w}_t, && (54)
\end{aligned}$$

where $\boldsymbol{\xi}_t$ is a vector of unobserved coefficients (states), $\mathbf{y}_t$ is a vector of observable variables, $\mathbf{x}_t$ is a vector of exogenous or predetermined variables, $\mathbf{F}(\mathbf{x}_t)$ and $\mathbf{H}(\mathbf{x}_t)$ are matrix-valued functions of $\mathbf{x}_t$, and $\mathbf{a}(\mathbf{x}_t)$ is a vector-valued function $\mathbf{x}_t$, all with conformable dimensions. The vectors of noises $\mathbf{v}_{t+1}$ and $\mathbf{w}_t'$ are assumed to be mutually independent and distributed according to a Gaussian distribution, conditionally on $\mathbf{I}_t = (\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1)$, with mean zero and variances given by $\mathbf{Q}(\mathbf{x}_t)$ and $\mathbf{R}(\mathbf{x}_t)$, respectively. Assuming further that the initial state $\boldsymbol{\xi}_1 \sim N\left(\hat{\boldsymbol{\xi}}_{1|0}, \mathbf{P}_{1|0}\right)$, the optimal estimates

of states $\boldsymbol{\xi}_t$ are obtained through the Kalman filter equations given by

$$\hat{\boldsymbol{\xi}}_{t|t} = \hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{K}_t \left[\mathbf{y}_t - \mathbf{a}(\mathbf{x}_t) - [\mathbf{H}(\mathbf{x}_t)]' \hat{\boldsymbol{\xi}}_{t|t-1}\right], \tag{55}$$

$$\mathbf{P}_{t|t} = \left[\mathbf{I} - \mathbf{K}_t [\mathbf{H}(\mathbf{x}_t)]'\right] \mathbf{P}_{t|t-1}, \tag{56}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) \left[[\mathbf{H}(\mathbf{x}_t)]' \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) + \mathbf{R}(\mathbf{x}_t)\right]^{-1}, \tag{57}$$

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}(\mathbf{x}_t) \hat{\boldsymbol{\xi}}_{t|t}, \tag{58}$$

$$\mathbf{P}_{t+1|t} = \mathbf{F}(\mathbf{x}_t) \mathbf{P}_{t|t} [\mathbf{F}(\mathbf{x}_t)]' + \mathbf{Q}(\mathbf{x}_t). \tag{59}$$

where the subscripts indicate the timing of information associated to each estimate, e.g., $t+1|t$ means the inference standing for period $t+1$ of the associated variable is made on the basis of data observed through period t.

To show that (20)-(22) represents the Kalman solution to the estimation of the time-varying coefficients of the model in (1) and (19), first let $\boldsymbol{\xi}_t \equiv \boldsymbol{\theta}_{j,t}$, $\mathbf{F}(\mathbf{x}_t) \equiv \mathbf{I}$, $\mathbf{v}_t \equiv \boldsymbol{\omega}_{j,t}$, $\mathbf{y}_t \equiv y_{j,t}$, $\mathbf{a}(\mathbf{x}_t) \equiv \mathbf{0}$, $\mathbf{H}(\mathbf{x}_t) \equiv \mathbf{x}_t$, $\mathbf{w}_t \equiv \varepsilon_{j,t}$, $\mathbf{Q}(\mathbf{x}_t) \equiv \boldsymbol{\Omega}_{j,t}$, and $\mathbf{R}(\mathbf{x}_t) \equiv \sigma^2_{j,t}$. Substituting these in (55)-(59) we get

$$\hat{\boldsymbol{\theta}}_{j,t|t} = \hat{\boldsymbol{\theta}}_{j,t|t-1} + \mathbf{K}_{j,t} \left[y_{j,t} - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{j,t|t-1}\right], \tag{60}$$

$$\mathbf{P}_{j,t|t} = [\mathbf{I} - \mathbf{K}_{j,t}\mathbf{x}_t'] \mathbf{P}_{j,t|t-1}, \tag{61}$$

$$\mathbf{K}_{j,t} = \mathbf{P}_{j,t|t-1}\mathbf{x}_t \left[\mathbf{x}_t'\mathbf{P}_{j,t|t-1}\mathbf{x}_t + \sigma^2_{j,t}\right]^{-1}, \tag{62}$$

$$\hat{\boldsymbol{\theta}}_{j,t+1|t} = \hat{\boldsymbol{\theta}}_{j,t|t}, \tag{63}$$

$$\mathbf{P}_{j,t+1|t} = \mathbf{P}_{j,t|t} + \boldsymbol{\Omega}_{j,t}. \tag{64}$$

Substituting (60) and (61) into (63) and (64) we get

$$\hat{\boldsymbol{\theta}}_{j,t+1|t} = \hat{\boldsymbol{\theta}}_{j,t|t-1} + \mathbf{K}_{j,t} \left[y_{j,t} - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{j,t|t-1}\right], \tag{65}$$

$$\mathbf{P}_{j,t+1|t} = [\mathbf{I} - \mathbf{K}_{j,t}\mathbf{x}_t'] \mathbf{P}_{j,t|t-1} + \boldsymbol{\Omega}_{j,t}, \tag{66}$$

which is evidently equivalent to the recursions in (20)-(22) with $\hat{\boldsymbol{\theta}}_{j,t} \equiv \hat{\boldsymbol{\theta}}_{j,t+1|t}$ and $\mathbf{P}_{j,t} \equiv \mathbf{P}_{j,t+1|t}$.

A.4 Correspondence between Kalman-based LS and *ad hoc* LS

Lets begin by rearranging terms in (3) and using the matrix inversion lemma,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}, \tag{67}$$

to find that

$$\gamma_{j,t}^{-1}\mathbf{R}_{j,t} = \underbrace{\gamma_{j,t}^{-1}\left(1-\gamma_{j,t}\right)\mathbf{R}_{j,t-1}}_{\mathbf{A}} + \underbrace{\mathbf{x}_t}_{\mathbf{U}}\overbrace{1}^{\mathbf{C}}\underbrace{\mathbf{x}_t'}_{\mathbf{V}}, \tag{68}$$

$$\gamma_{j,t}\mathbf{R}_{j,t}^{-1} = \frac{\gamma_{j,t}}{1-\gamma_{j,t}}\mathbf{R}_{j,t-1}^{-1} - \frac{\gamma_{j,t}}{1-\gamma_{j,t}}\mathbf{R}_{j,t-1}^{-1}\mathbf{x}_t\left(1+\mathbf{x}_t'\frac{\gamma_{j,t}}{1-\gamma_{j,t}}\mathbf{R}_{j,t-1}^{-1}\mathbf{x}_t\right)^{-1}\mathbf{x}_t'\frac{\gamma_{j,t}}{1-\gamma_{j,t}}\mathbf{R}_{j,t-1}^{-1}, \tag{69}$$

$$\frac{\gamma_{j,t}}{\gamma_{j,t-1}}\mathbf{R}_{j,t}^{-1} = \frac{\gamma_{j,t}}{\gamma_{j,t-1}\left(1-\gamma_{j,t}\right)}\left(\mathbf{I}-\frac{\mathbf{R}_{j,t-1}^{-1}\mathbf{x}_t\mathbf{x}_t'}{\frac{1-\gamma_{j,t}}{\gamma_{j,t}}+\mathbf{x}_t'\mathbf{R}_{j,t-1}^{-1}\mathbf{x}_t}\right)\mathbf{R}_{j,t-1}^{-1}, \tag{70}$$

$$\mathbf{P}_{j,t}^{LS} = \frac{\gamma_{j,t}}{\gamma_{j,t-1}\left(1-\gamma_{j,t}\right)}\left(\mathbf{I}-\frac{\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t+\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}\right)\mathbf{P}_{j,t-1}^{LS}, \tag{71}$$

where in the last line we let $\mathbf{P}_{j,t}^{LS} \equiv \gamma_{j,t}\mathbf{R}_{j,t}^{-1} \Rightarrow \mathbf{P}_{j,t-1}^{LS} = \gamma_{j,t-1}\mathbf{R}_{j,t-1}^{-1}$.

For the coefficients estimates recursion in (2), let $\mathbf{K}_{j,t}^{LS} \equiv \gamma_j\mathbf{R}_{j,t}^{-1}\mathbf{x}_t = \mathbf{P}_{j,t}^{LS}\mathbf{x}_t$ to then obtain

$$\mathbf{K}_{j,t}^{LS} = \frac{\gamma_{j,t}}{\gamma_{j,t-1}\left(1-\gamma_{j,t}\right)}\left(\mathbf{I}-\frac{\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t+\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}\right)\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t, \tag{72}$$

$$= \frac{\gamma_{j,t}}{\gamma_{j,t-1}\left(1-\gamma_{j,t}\right)}\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t\left(1-\frac{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t+\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}\right), \tag{73}$$

$$= \frac{\gamma_{j,t}}{\gamma_{j,t-1}\left(1-\gamma_{j,t}\right)}\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t\left(\frac{\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t+\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}\right), \tag{74}$$

$$\mathbf{K}_{j,t}^{LS} = \frac{\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{j,t-1}^{LS}\mathbf{x}_t+\frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1-\gamma_{j,t}\right)}. \tag{75}$$

Note that (75) and (71) assume exactly the same form as of (26) and (27), respectively, thus establishing the correspondence at scrutiny.

A.5 Correspondence between Kalman filter and SG

First, lets show how (28) is obtained from substitution of (11) into (23). To see that, first notice that an scalar λ_t equivalent to $\boldsymbol{\lambda}_t$ in (11) can be obtained by solving the eigenvalue problem

$$\left(\mathbf{I}-\mu_t\mathbf{x}_t\mathbf{x}_t'\right)\mathbf{z} = \lambda_t\mathbf{z}, \tag{76}$$

for an arbitrary $K\times 1$ vector $\mathbf{z}$. As usual, this problem can be solved by finding the λ_t that makes the following determinant to be equal to zero,

$$\det\left[\left(1-\lambda_t\right)\mathbf{I}-\mu_t\mathbf{x}_t\mathbf{x}_t'\right] = 0. \tag{77}$$

Making use of the matrix determinant lemma, i.e.,

$$\det\left[\mathbf{A}+\mathbf{u}\mathbf{v}'\right] = \det\left[\mathbf{A}\right]\left(1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}\right), \tag{78}$$

we then find that

$$\det\left[\underbrace{(1-\lambda_t)\,\mathbf{I}}_{\mathbf{A}}\underbrace{-\mu_t\mathbf{x}_t}_{\mathbf{u}}\underbrace{\mathbf{x}_t'}_{\mathbf{v}'}\right] = (1-\lambda_t)^K\left(1-\mathbf{x}_t'\,(1-\lambda_t)^{-1}\,\mathbf{I}\mu_t\mathbf{x}_t\right), \tag{79}$$

$$= (1-\lambda_t)^K - (1-\lambda_t)^{K-1}\,\mu_t\mathbf{x}_t'\mathbf{x}_t. \tag{80}$$

Equating this last expression to zero and solving then we find that

$$\lambda_t = 1-\mu_t\mathbf{x}_t'\mathbf{x}_t, \tag{81}$$

which is the scalar version of $\boldsymbol{\lambda}_t$ in (11) that we were looking for substitution into (23), together with α_t from (11) as well. This results in

$$\sigma_{j,t}^2 = (1-\mu_t\mathbf{x}_t'\mathbf{x}_t)\,\mu_t^{-1}, \tag{82}$$

$$= \mu_t^{-1} - \mathbf{x}_t'\mathbf{x}_t, \tag{83}$$

which is recognizably equal to (28) as we wanted to show.

Now, to show the correspondence between the Kalman filter and the SG algorithm, start by substituting (28) into the Kalman gain formulae, (21), to find that

$$\mathbf{K}_{j,t} = \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \mu_t^{-1} - \mathbf{x}_t'\mathbf{x}_t}, \tag{84}$$

$$= \frac{\mu_t\mathbf{P}_{j,t-1}\mathbf{x}_t}{\mu_t\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t - \mu_t\mathbf{x}_t'\mathbf{x}_t + 1}, \tag{85}$$

where clearly we need $\mathbf{P}_{j,t-1} = \mathbf{I}$ in order to (30) hold true. To achieve this, we can use (29) into (22) to find that

$$\mathbf{P}_{j,t} = \left(\mathbf{I} - \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma_{j,t}^2}\right)\mathbf{P}_{j,t-1} + \mathbf{I} - \left(\mathbf{I} - \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma_{j,t}^2}\right)\mathbf{P}_{j,t-1}, \tag{86}$$

$$\mathbf{P}_{j,t}^{SG} = \mathbf{I}. \tag{87}$$

Substituting this result into the previous formulae for the Kalman gain specialized to the SG case we find

$$\mathbf{K}_{j,t}^{SG} = \mu_t\mathbf{x}_t, \tag{88}$$

thus confirming the correspondence between the Kalman filter and the SG algorithm under the assumptions of (28) and (29).