Learning can generate Long Memory[*]

Guillaume Chevillon[†]
ESSEC Business School & CREST

Sophocles Mavroeidis[‡]
University of Oxford

December 1, 2016

Abstract

We study learning dynamics in a prototypical representative-agent forward-looking model in which agents' beliefs are updated using linear learning algorithms. We show that learning in this model can generate long memory endogenously, without any persistence in the exogenous shocks, depending on the weights agents place on past observations when they update their beliefs, and on the magnitude of the feedback from expectations to the endogenous variable. This is distinctly different from the case of rational expectations, where the memory of the endogenous variable is determined exogenously.

JEL Codes: C1, E3;

Keywords: Long Memory, Recursive Least Squares, Decreasing Gain Learning, New Keynesian Phillips curve.

1   Introduction

In many economic models, the behavior of economic agents depends on their expectations of the current or future states of the economy. For example, in forward-looking New Keynesian models, prices are set according to firms' expectations of future marginal costs, consumption is determined according to consumers' expectations of future income, and policy makers' actions depend on their expectations of the current and future macroeconomic conditions, see Clarida, Galí and Gertler (1999). In asset pricing models, prices are determined by expected dividends and future price appreciation, see Campbell and Shiller (1987).

In a rational expectations equilibrium, these models imply that the dynamics of the endogenous variables are determined exogenously and therefore, these models typically fail to explain the observed persistence in the data. It has long been recognized that bounded rationality, or learning, may induce richer dynamics and can account for some of the persistence in the data, see Sargent (1993) and Evans and Honkapohja (2009).

The objective of this paper is to point out the connection between learning and a specific form of persistence, namely long memory. In particular, we show that in certain economic models, replacing rational expectations with recursive least squares learning can generate long memory, i.e. stronger persistence than usually considered in the learning literature (e.g. Milani, 2007, Eusepi and Preston, 2011, or Adam, Marcet and Nicolini, 2016). We focus on a prototypical representative-agent forward-looking model and least-squares learning algorithms, which are popular in theoretical and empirical

work, see Evans and Honkapohja (2009). This framework is simple enough to obtain analytical results, but sufficiently rich to nest several interesting applications. We find that the incidence and extent of the long memory depends both on how much agents discount past observations when updating their beliefs, and on the magnitude of the feedback that expectations have on the process. The latter is governed by the coefficient on expectations, which in many applications is interpretable as a discount factor. It is important to stress that this coefficient plays no role for the memory of the process under rational expectations. These results are established under the assumption that exogenous shocks have short memory, and hence, it is shown that long memory can arise completely endogenously through learning. Finally, we provide a short empirical illustration on the New Keynesian Phillips curve.

The above results provide a new structural interpretation of a phenomenon which has been found to be important for many economic time series. The other main explanations of long memory that we are aware of are: (i) aggregation of short memory series — either cross-sectionally (with beta-distributed weights in Granger, 1980, or with heterogeneity in Abadir and Talmain, 2002, and Zaffaroni, 2004), or temporally across mixed-frequencies (Chambers, 1998); (ii) occasional breaks that can produce fractional integration (Parke, 1999) or be mistaken for it (Granger and Ding, 1996, Diebold and Inoue, 2001, or Perron and Qu, 2007); (iii) some form of nonlinearity (see, *e.g.,* Davidson and Sibbertsen, 2005, and Miller and Park, 2010); or (iv) within large systems, with reinforcement through networks (Schennach, 2013) or via marginalization of large dimensional processes (Chevillon, Hecq and Laurent, 2015). Ours is the first

explanation that traces the source of long memory to the behavior of agents, and the self-referential nature of economic outcomes. We would like to stress that the models we study here do not yield fractional integration. In that sense, the long memory that is induced by learning can be referred to as apparent (Davidson and Sibbertsen, 2005) or spurious (Haldrup and Kruse, 2014).

The paper is organized as follows. Section 2 introduces the modelling framework and characterization of learning algorithms. We then present in Section 3 our analytical results. Monte Carlo simulation evidence confirming our theoretical predictions follows in Section 4. Finally, in Section 5 we present an empirical illustration. Proofs are given in the Appendix at the end. Supplementary material collecting further proofs and simulation results is available online.

Throughout the paper, $f(x) \sim g(x)$ as $x \to a$ means $\lim_{x \to a} f(x) / g(x) = 1$. We also use the Bachmann-Landau notations to denote orders of magnitude: $f(x) = O(g(x))$ means there exist $M > 0$ such that $|f(x)| \leq M |g(x)|$ in a neighborhood of a; $f(x) = o(g(x))$ means $f(x) / g(x) \to 0$; and $f(x) \asymp g(x)$ denotes the "exact rate", i.e., $f(x) = O(g(x))$ and $g(x) = O(f(x))$. Also, we use the notation $\mathsf{sd}(X)$ to refer to the standard deviation $\sqrt{\mathsf{Var}(X)}$.

2   Framework

We consider a linear representative agent framework with constant parameters, so as to avoid confounding our results with other well-known sources of long-range de-

pendence discussed above. For example, the representative agent assumption avoids inducing long memory through heterogeneity and aggregation, as in, *e.g.,* Granger (1980), Abadir and Talmain (2002), Zaffaroni (2004) and Schennach (2013).

We study two different forward-looking models that link an endogenous variable y_t to an exogenous process x_t. The first model is:

$$y_t = \beta y_{t+1}^e + x_t, \quad t = 1, 2, ..., T, \tag{1}$$

where y_{t+1}^e denotes the expectation of y_{t+1} conditional on information up to time t. This model features one-step-ahead forecasts of the endogenous variable y_t, and has been studied extensively in the learning literature, see Evans and Honkapohja (2001). The second model involves infinite-horizon expectations of x_t and is given by

$$y_t = x_t + \sum_{j=1}^{\infty} \beta^j x_{t+j}^e, \quad t = 1, 2, ..., T, \tag{2}$$

provided that the infinite sum converges. Under rational expectations, we set $y_{t+1}^e = E_t(y_{t+1})$ and $x_{t+j}^e = E_t(x_{t+j})$ where E_t denotes expectations based on the true law of motion of y_t and x_t, respectively. When $|\beta| < 1$ and there are no bubbles ($\lim_{T \to \infty} |E_t(y_T)| < \infty$ for all $t < \infty$), the specifications (1) and (2) are equivalent (see e.g. Blanchard and Fischer, 1989, sec. 5.1), but under adaptive learning, they are not (see Preston, 2005, 2006). So, we will analyze each of these specifications separately below.

Under adaptive learning (Sargent, 1993, Evans and Honkapohja, 2001), agents in the model are assumed to face the same limitations as empirical economists, who postulate models but "do not know their parameter values and must estimate them

econometrically". Thus, agents are assumed to "act like statisticians or econometricians when doing the forecasting about the future state of the economy" (Evans and Honkapohja, 2001, p. 13).[1] Specifically, agents form expectations based on some perceived law of motion (PLM) for the process y_t or x_t, whose parameters are recursively estimated using information available to them. Their forecasts or learning algorithms can be expressed as weighted averages of past data, where the weights may vary over time to reflect information accrual as the sample increases, which is a key aspect of learning.

For the objectives of this paper, it is important to restrict attention to linear learning algorithms, so as to emphasize that long range dependence can arise without the need for nonlinearities – contrast this with Diebold and Inoue (2001), Davidson and Sibbertsen (2005) and Miller and Park (2010) (see also the surveys by Granger and Ding, 1996, and Davidson and Teräsvirta, 2002). Linear learning algorithms can be motivated using the so-called mean-plus-noise model as PLM, see equations (3) and (4) below. It is typical in the literature to assume that the PLM nests some rational expectations equilibrium, so that agents in the model are in some sense 'boundedly rational' (Sargent, 1993). This is not essential for our results, but we will maintain that convention here.[2] To do so, we use as our baseline assumption that x_t is indepen-

---

[1]The literature on learning in economics is vast and spans several decades. For extensive justification of the learning approach to the modelling of expectations, see the monographs by Sargent (1993) and Evans and Honkapohja (2001), and the references therein.

[2]Note that it is still possible to justify the use of the mean-plus-noise PLM even when it does not nest the rational expectations equilibrium, using the concept of 'restricted perceptions equilibrium'

dently and identically distributed (*i.i.d.*), but note that the results of the paper also hold when x_t is persistent but has short memory, see Assumption A' below.

Assumption A. The process x_t is determined by

$$x_t = \mu + \epsilon_t, \tag{3}$$

where μ is a constant and ϵ_t is *i.i.d.*, with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) < \infty$.

When Assumption A holds and $|\beta| < 1$, the unique stationary solution of (1) under rational expectations is

$$y_t = \alpha + \epsilon_t, \tag{4}$$

where $\alpha = \mu/(1-\beta)$, see, e.g., Evans and Honkapohja (2001, ch. 8).[3] We assume that in model (1), agents forecast y_{t+1} using the PLM (4). Under this PLM, the conditional expectation of y_{t+1} given information up to time t is simply α, and because it is unknown to the agents, their forecast y_{t+1}^e is given by a recursive estimate of α. The classic learning algorithm in the literature is recursive least squares (RLS): $y_{t+1}^e = \frac{1}{t} \sum_{i=1}^{t} y_i$. This formula implicitly assumes that the information set of the agents starts at the same time as our observed sample. One could look for plausible justifications for this, but it is best to view this assumption as an inconsequential simplification, because the results of this paper on the long-range dependence induced by learning dynamics continue to hold if agents' sample is assumed to start at any fixed

(Evans and Honkapohja , 2001, sec. 3.6), or 'self-confirming equilibrium' (Sargent, 1999).

[3]This is also known as the minimum state variable solution of McCallum (1983) and also frequently referred as the 'fundamental solution' (Evans and Honkappohja, 2001, p. 178).

point $-\infty < t_0 \leq 1.$[4] So, for simplicity and following the convention in the learning literature, we will assume that agents' sample starts at $t = 1$.

RLS is a member of the class of weighted least squares (WLS) algorithms (Sargent, 1993) that are defined as the solution to the minimization problem

$$y_{t+1}^e = \underset{a}{\arg\min} \sum_{j=0}^{t-1} w_{t,j} (y_{t-j} - a)^2, \quad \sum_{j=0}^{t-1} w_{t,j} = 1. \tag{5}$$

RLS corresponds to $w_{t,j} = t^{-1} 1_{\{0 \leq j < t\}}$ where $1_{\{.\}}$ denotes the indicator function.[5] Another member of this class, which is particularly popular in applied work, obtains when the weights decline exponentially, i.e., $w_{t,j} \propto (1 - \bar{g})^j$ for some constant $\bar{g} \in (0, 1)$.

The above algorithms can also be expressed alternatively as stochastic recursive algorithms of the form

$$a_t = a_{t-1} + g_t (y_t - a_{t-1}), \quad t \geq 1, \tag{6}$$

where a_0 is a parameter that characterizes initial beliefs, and the sequence $\{g_t\}$ is known as the "gain" (see Evans and Honkapohja, 2001, ch. 6). The literature distinguishes between decreasing gain algorithms, which have the property that $g_t \to 0$ as $t \to \infty$, and constant gain algorithms, with g_t equal or converging to some $\bar{g} > 0$. RLS obtains by setting $g_t = 1/t$, and is a member of a more general class of decreasing gain least squares (DGLS) algorithms where $g_t \sim \theta t^{-\nu}$, with $\theta > 0$ and $\nu \in (0, 1]$,

---

[4]The limiting case $t_0 \to -\infty$ is not interesting because it has no learning dynamics. Specifically, when $\beta < 1$, $y_{t+1}^e \to \alpha$ with probability one as $t_0 \to -\infty$, so expectations become rational, see Evans and Honkapohja (2001, ch. 8).

[5]It should be noted that equation (5) allows expectations at date t to depend on the realization at date t of y_t, meaning that y_{t+1}^e and y_t are simultaneously determined.

as discussed in Evans and Honkapohja (2001, ch. 7). Malmendier and Nagel (2016) recently considered an application where $\nu = 1$ and θ is interpreted as a "forgetting factor", in the terminology of Marcet and Sargent (1989) who consider a related algorithm. This algorithm belongs to the class of weighted least squares, see Section 1 in the Supplement for details.

The most typical constant gain algorithm is constant gain least squares (CGLS), and obtains by setting $g_t = \bar{g}$ in (6). This is the same as WLS (5) with $w_{t,j} = \bar{g}\left(1 - \bar{g}\right)$, so agents discount past information exponentially fast when forming their forecasts. As we discuss in the next section, CGLS does not generate long memory in the models that we consider, so we will focus our discussion in this paper on decreasing gain algorithms only.[6]

For model (2), agents obtain infinite-horizon forecasts of x_{t+j}, $j \geq 1$, conditional on information at time t, using (3) as their PLM. Under this PLM, the conditional expectation of x_{t+j} given information up to time t is simply μ, which agents learn using a learning algorithm analogous to (6). Denoting their recursive estimates of μ by b_t, we have

$$b_t = b_{t-1} + g_t\left(x_t - b_{t-1}\right), \quad t \geq 1, \tag{7}$$

and b_0 is a parameter that characterizes initial beliefs. Hence, agents' forecasts under learning are given by $x_{t+j}^e = b_t$ for all $j \geq 1$.

Finally, we need to specify a working definition of long memory or long-range dependence. There are several measures of dependence that can be used to characterize

---

[6]See Chevillon and Mavroeidis (2017) for results under CGLS learning.

the memory of a stochastic process, such as mixing coefficients and autocorrelations (when they exist). Various alternative definitions of short memory are available (*e.g.*, various mixing conditions, see White, 2000). These definitions are not equivalent, but they typically imply that short memory requires that the variance of partial sums, scaled by the sample size, T, should be bounded.[7] If this does not hold, we will say that the process exhibits long memory.[8] We can also define the 'degree of memory' of a second-order process z_t (i.e. such that $E\left(z_t^2\right) < \infty$ for all t) by the parameter d (when it exists) such that

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} z_t\right) \asymp T^d. \tag{8}$$

Definition LM (long memory) *A second-order z_t exhibits long memory of degree $d > 0$ if there exists $d > 0$ such that (8) holds. The process z_t exhibits short memory if $d = 0$.*

The above definition applies generally to any stochastic process that has finite second moments (which we assume in this paper). For example, Definition LM is not confined to fractionally integrated processes, and it allows for processes that are non-

---

[7]Any definition of short memory that implies an invariance principle satisfies the restriction on the variance of partial sums, e.g., Andrews and Pollard (1994), Rosenblatt (1956), or White (2000).

[8]This is also the definition adopted by Diebold and Inoue (2001) in their study of the connection between structural change and long memory. An alternative measure of long memory is the total variation distance from a process that is integrated of order zero (I(0)), see Müller and Watson (2008). Unfortunately, this does not seem to be analytically tractable in the models that we study here.

stationary even when $d \in (0, 1/2)$. This is important in the present context because the models that we study here are inherently nonstationary, so we cannot use alternative definitions of long memory that require stationarity.[9]

3 Analytical results

This section provides our main results. We start by showing that, in the model that we consider, long memory cannot arise endogenously under rational expectations. We then analyze the impact of learning on the memory of the resulting process.

3.1 Memory under Rational Expectations

The model with rational expectations is given by

$$y_t = \beta E_t \left(y_{t+1} \right) + x_t, \quad t = 1, 2, ..., T. \tag{9}$$

When $|\beta| < 1$, the unique stationary solution of (9) is given by

$$y_t = \sum_{j=0}^{\infty} \beta^j E_t \left(x_{t+j} \right), \quad t = 1, 2, ..., T, \tag{10}$$

see Gourieroux et al. (1982).[10] It is trivial to see that under Assumption A, $E_t \left(x_{t+j} \right) = \mu$ for all $j > 0$ and the rational expectations solution (10) can be written as (4), so y_t exhibits short memory. It is nontrivial to show that the same result applies even

---

[9]For a covariance stationary process, there are alternative equivalent definitions of long memory based on the decay of the autocorrelation function and the spectrum, see Beran (1994).

[10]This is also known as the "fundamental solution", see Blanchard and Fisher (1989, sec. 5.1).

when x_t exhibits more persistence than under Assumption A. In particular, we wish to investigate whether under rational expectations, the endogenous variable y_t can exhibit long memory when the forcing variable x_t has short memory. To do this, we consider the following generalization of Assumption A to include typical covariance stationary processes with short memory.

Assumption A'. The process x_t is determined by $x_t = \mu + \sum_{j=0}^{\infty} \vartheta_j \epsilon_{t-j}$, where μ is a constant, ϵ_t is *i.i.d.* with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) < \infty$, and the series ϑ_j satisfies $\sum_{j=0}^{\infty} |\vartheta_j| < \infty$ and $\sum_{j=0}^{\infty} \vartheta_j \neq 0$.

The next proposition shows that in the class of models we consider, long memory cannot arise endogenously under rational expectations.

Proposition 1 *Suppose x_t satisfies Assumption A', and let y_t be given by equation (10) with $|\beta| < 1$. Then,* $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) = O(1)$.

This result suffices to establish that y_t cannot satisfy Definition LM with $d > 0$, so it cannot exhibit long memory by that definition.[11] Proposition 1 also shows that the magnitude of the feedback that expectations have on the process, which is measured by β, plays no role for the memory of the process under rational expectations. As we

---

[11]Note that Proposition 1 does not establish an exact rate, $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) \asymp 1$, because it is not possible to rule out the case of antipersistence, which would arise if $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) = o(1)$, as the following example demonstrates. Suppose $x_t = \mu + \epsilon_t - (1+\beta)^{-1}\epsilon_{t-1}$. Then $y_t = \sum_{j=0}^{\infty} \beta^j E_t(x_{t+j})$ $= \mu(1+\beta) + \left(1 - \frac{\beta}{1+\beta}\right)\epsilon_t - \frac{1}{1+\beta}\epsilon_{t-1} = \mu(1+\beta) + \frac{1}{1+\beta}\Delta\epsilon_t$.

will see below, this is very different from what happens under decreasing gain learning in model (1), where the memory of y_t crucially depends on the proximity of β to 1.

3.2 Memory under learning

We now turn to the model with learning. We will focus on decreasing gain learning and comment on constant gain learning briefly at the end of this section. The analysis in this subsection relies on Assumption A, but the results in Theorems 2 and 3 given below hold more generally under Assumption A', see Appendix B for details.

The learning literature has focused mainly on the convergence properties of decreasing gain learning algorithms to rational expectations. In model (1) and under Assumption A, DGLS learning converges almost surely to rational expectations, that is, $\Pr\left(\lim_{t\to\infty} y_{t+1}^e = \alpha\right) = 1$, see, *e.g.*, Evans and Honkapohja (2001, Theorem 6.10 and ch. 8). However, the literature has not studied the effect of transitional dynamics on the memory of y_t. This is the purpose of the following theorem.

Theorem 2 *Consider the model (1), with $\beta < 1$, $\beta \neq 0$, $y_{t+1}^e = a_t$ as given in equation (6) where $g_t = \theta/t + O\left(t^{-1-\zeta}\right)$, $\zeta > 0$, $\theta > 0$, $E\left(a_0^2\right) < \infty$ and x_t satisfies Assumption A. Then, as $T \to \infty$,*

$$
\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) \asymp \begin{cases} T^{\frac{1}{2}-\theta(1-\beta)}, & \text{if } \theta\left(1-\beta\right) < \frac{1}{2}, \\[2mm] \sqrt{\log T}, & \text{if } \theta\left(1-\beta\right) = \frac{1}{2}, \\[2mm] 1, & \text{if } \theta\left(1-\beta\right) > \frac{1}{2}. \end{cases}
$$

The theorem shows that the process y_t exhibits long memory of degree $d \in \left(0, \frac{1}{2}\right)$

13

when $\beta > 1 - \frac{1}{2\theta}$. The degree of memory is $\max\left(1/2 - \theta\left(1 - \beta\right), 0\right)$. For RLS $(\theta = 1)$ this specializes to $\max\left(\beta - \frac{1}{2}, 0\right)$.

To gain some intuition for this result, consider the impulse response function (IRF) of y_{t+j} with respect to x_t in the model used in Theorem 2 with $g_t = \min\left(1, \theta/t\right)$. From the moving average representation of y_{t+j} derived in the Supplement, Expression (S-2), we obtain, for $t \geq \theta$ and as $j \to \infty$,

$$\frac{\partial y_{t+j}}{\partial x_t} \sim \beta\theta \frac{\Gamma\left(t - \beta\theta\right)}{\Gamma\left(t + 1 - \theta\right)} j^{-\theta(1-\beta)}. \tag{11}$$

Expression (11) shows that the IRF decays hyperbolically in j. Moreover, the closer β is to unity, or the smaller is the forgetting factor θ, the slower is the decay of the response with respect to the horizon j.

Note that the long memory we obtain here is clearly not of the type of a stationary fractionally integrated process. Therefore, the limit theory that applies to fractionally integrated processes, such as the convergence of normalized partial sums to fractional Brownian motion, does not necessarily apply in the context treated in this paper. If fractionality is required in the definition of long memory, then the long memory that is generated by learning could be considered of a spurious variety. However, it is important to note that the long memory generated by learning is not of the same type as the spurious long memory induced by structural breaks whose frequency is linked to the observed sample (Diebold and Inoue, 2001, Perron and Qu, 2009), which has been studied in many recent papers, see Haldrup and Kruse (2014) and the references therein. The key difference is that our data generating process is not a triangular array,

14

i.e., the parameters of the model are not linked to the size of the sample observed by the econometrician. Moreover, we have reasons to believe that tests designed against spurious long memory induced by structural change may not have much power against long memory induced by learning.[12]

The theorem explains a result from the learning literature on the properties of agents' forecasts under decreasing gain learning: even though y^e_{t+1} converges to the rational expectation $E_t(y_{t+1}) = \alpha$ when $\beta < 1$, asymptotic normality of y^e_{t+1} is only established when $\beta < 1 - \frac{1}{2\theta}$ (Marcet and Sargent, 1995, and Evans and Honkapohja, 2001, Theorem 7.10). Moreover, Marcet and Sargent (1995) conjectured, based on simulations, that learning converges to rational expectations more slowly than at rate $\sqrt{t}$ when $\beta > 1 - \frac{1}{2\theta}$ (in the model we study here). Theorem 2 explains why standard central limit theory does not apply to agents' estimator y^e_{t+1} when $\beta \geq 1 - \frac{1}{2\theta}$. By corollary, the rate at which y^e_{t+1} converges to the rational expectation is then slower than $\sqrt{t}$, which proves Marcet and Sargent's (1995) conjecture for this model.

Theorem 2 covered the case of DGLS learning in the model of equation (1). The next theorem gives the corresponding result for the model with infinite horizon learning, given in equation (2).

---

[12]One such test by Haldrup and Kruse (2014) is based on the idea that polynomial transformations of a fractionally integrated process have a different degree of integration than the same transformations applied to spurious long memory processes. In the supplement, we show that under Gaussian errors, the degree d_2 of long memory of y_t^2 relates to d, the memory of y_t, in exactly the same way as if y_t were fractionally integrated, i.e., $d_2 = \max\{0, 2d - 1/2\}$. This is different from what Haldrup and Kruse (2014, sec. 3) report for the process of Diebold and Inoue (2001) to motivate their test.

Theorem 3 *Consider the model (2) with* $\beta < 1$, $\beta \neq 0$, $x^e_{t+j} = b_t$ *as given in equation (7) where* $g_t = \theta/t + O\left(t^{-1-\varsigma}\right)$, $\varsigma > 0$, $\theta > 0$, $E\left(b_0^2\right) < \infty$ *and* x_t *satisfies Assumption A. Then, as* $T \to \infty$,

$$\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) \asymp \begin{cases} T^{\frac{1}{2}-\theta}, & \text{if } \theta < \frac{1}{2}, \\[2mm] \sqrt{\log T}, & \text{if } \theta = \frac{1}{2}, \\[2mm] 1, & \text{if } \theta > \frac{1}{2}. \end{cases}$$

The above result shows that the process exhibits long memory of degree $d \in \left(0, \frac{1}{2}\right)$ when $\theta \in \left(0, \frac{1}{2}\right)$. The degree of memory is $\max\left(1/2 - \theta, 0\right)$. It is not surprising that this result is very similar to Theorem 2, because inspection of the proof of both theorems in the Appendix reveals that both models admit essentially the same moving average representation. The important difference, however, is that the degree of memory induced by the so-called 'infinite-horizon learning' in model (2) does not depend on β, the coefficient on future expectations. In infinite-horizon learning, memory only depends on the forgetting factor θ, which is inversely related to the weight that the learning algorithm puts on distant observations, and long memory can arise only if the learning algorithm has a sufficiently low forgetting factor. For example, long memory cannot arise under RLS in infinite-horizon learning, whereas it can arise under RLS in model (1). The reason for this difference is that in the infinite-horizon specification, agents are learning about the exogenous process x_t, and there is no feedback from the learning algorithm to the future values of the process the agents are learning about, since it is exogenous. In contrast, in model (1), agents are learning about the endogenous process y_t, and learning affects its law of motion. This "self-referential" nature

16

of learning dynamics (Evans and Honkapohja, 2001) in (1) appears to be the cause of the stronger persistence relative to infinite horizon learning.

Finally, we conclude this section by noting that long memory does not arise under constant gain least-squares algorithms, known as 'perpetual learning'. In a companion paper, Chevillon and Mavroeidis (2017), we show that with perpetual learning the ALM corresponding to model (1) admits a stationary representation with absolutely summable autocovariances when $|\beta| < 1$. In this case, apparent long memory arises when β is modelled as local to unity, i.e., when the process y_t is modelled as a triangular array.

4   Simulations

This section presents simulation evidence in support of the analytical results given above. We generate samples of $\{y_t\}$ from (1) under the RLS learning algorithm. The exogenous variable x_t is assumed to be *i.i.d.* normal with mean zero, and its variance is normalized to 1 without loss of generality. We use a relatively long sample of size $T = 1000$ and various values of the parameter β. We study the behavior of the variance of partial sums, the spectral density, and the popular Geweke and Porter-Hudak (1983, henceforth GPH) and the Robinson (1995) maximum local Whittle likelihood (labelled Whittle in the tables) estimators of the long memory parameter d.[13] We also report the power of tests of the null hypotheses $d = 0$ and $d = 1$, as the empirical rejection

---

[13]We use $n = \lfloor T^{1/2} \rfloor$ Fourier ordinates, where $\lfloor x \rfloor$ denote the integer part of x.

frequencies of one-sided 5% level tests of $\mathsf{H}_0 : d = 0$ against $\mathsf{H}_1 : d > 0$, and $\mathsf{H}_0 : d = 1$ against $\mathsf{H}_1 : d < 1$, resp. The number of Monte Carlo replications is 10,000. Additional figures reporting the rate of growth of the variance of partial sums are available in the Supplement.

Figure 1 reports the Monte Carlo average log sample periodogram against the log frequency ($\log \omega$). This constitutes a standard visual evaluation of the presence of long range dependence if the log periodogram is linearly decreasing in $\log \omega$. The figure indicates that y_t exhibits long memory for $\beta > 1/2$ and the degree of long memory increases with β. Table 1 records the means of the estimators, and the empirical rejection frequency (power) of tests of the hypotheses $d = 0$ and $d = 1$ (the latter is based on a test of $d = 0$ for Δy_t) against the one-sided alternatives $d > 0$ and $d < 1$ respectively. Evidently, $E(\widehat{d})$ increases with β in accordance with Theorem 2, *i.e.*, $E(\widehat{d}) \approx \max\left(0, \beta - 1/2\right)$.

Unreported figures (available in the Supplement) show that the log of $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right)$ increases linearly with $\log T$ and that the growth rate of the ratio $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) / \log T$ tends quickly to the values the theorems imply for the degree of memory under RLS learning.

5 Empirical illustration

This section provides a short empirical illustration of the above results on the low frequency variation of inflation using the New Keynesian Phillips curve (NKPC) where

18

expectations are formed using DGLS learning. The model corresponds to equation (1) or (2) with y_t denoting inflation and x_t being proportional to the output gap or marginal costs, see, e.g., Galí and Gertler (1999).

We use monthly U.S. consumer price index (CPI) inflation data obtained from the database of the Federal Reserve Bank of Philadelphia. The database contains a non-seasonally adjusted series over the period 1913(2)-2015(8), and a seasonally adjusted series over the period 1947(1)-2015(8). Table 2 reports the estimated degrees of long memory for these series, using the Two-Step Exact Whittle Likelihood estimator (2EWL) of Shimotsu and Phillips (2005) and Shimotsu (2010) with bandwidths set to the square roots of the sample sizes. All estimates are close to 0.4. These estimates correspond to results reported in Haldrup and Kruse (2014) who studied the long range dependence in those same series over a slightly shorter sample.

Next, we ask whether it is possible to explain the observed low frequency variation in inflation endogenously using learning, that is, when the exogenous process x_t exhibits short memory. In our empirical analysis, we calibrate the monthly discount factor β to 0.997, to correspond to an annualized discount rate of 4%, as is commonly assumed, see, e.g., Clarida et al. (1999). We study both (1) and (2) specifications for learning, referring to them as Euler Equation (EE) and Infinite Horizon (IH) specifications, following Preston (2006). We compute the expectation under learning, denoted $y_{t+1}^{e,EE}(\theta)$ for EE and $x_{t+1}^{e,IH}(\theta)$ for IH, where $y_{t+1}^{e,EE}(\theta) = a_t$ and $x_{t+1}^{e,IH}(\theta) = b_t$ are given by the DGLS algorithms (6) and (7), respectively, with $g_t = \min(1, \theta/t)$ and forgetting

factor θ.[14] Initial beliefs are set equal to the first observation in the sample. Next, we compute $x_t^{EE}(\theta) = y_t - \beta y_{t+1}^{e,EE}(\theta)$ and $x_t^{IH}(\theta) = y_t - \frac{\beta}{1-\beta} x_{t+1}^{e,IH}(\theta)$. We can then test whether there is any value of θ for which $x_t^{EE}(\theta)$ and $x_t^{IH}(\theta)$ exhibit short memory. For any given θ, we can do so using a standard t test derived from the 2EWL estimator and its asymptotic standard error. We can also obtain $(1 - \eta)$-level confidence intervals for θ in EE and IH by inverting two-sided η-level t tests, i.e., by collecting all the values of θ for which the test accepts the null hypothesis that the memory parameter of $x_t^{EE}(\theta)$ or $x_t^{IH}(\theta)$ is equal to zero, respectively. If this confidence interval is not empty, i.e., if there is a value of θ for which the test accepts the null hypothesis, we can conclude that there is a learning algorithm of the type indexed by θ that can explain the low frequency variation in y_t.

We consider values of $\theta \geq 0.01$ in steps of 0.01 over $[0.01, 1]$ and of unity for $\theta \geq 1$. Table 2 reports 90% confidence intervals based on the 2EWL t test for the two different specifications and the two different data sets. In the non-seasonally adjusted data set (1913-2015) the confidence intervals are $[0.03, 0.08]$ for IH and $[8, 27]$ for EE. Note that Theorems 2 and 3 imply that for any θ in the IH specification, the corresponding θ that yields the same long memory for y_t under EE is $(1 - \beta)^{-1}$ times larger. These

---

[14]An interesting generalization of the above DGLS learning algorithm is $g_t = \theta/(t + T_0)$, with $T_0 \geq 0$ an integer representing agents' prior estimation sample. As remarked earlier, this generalization does not affect the theoretical results of the paper, but it could make a difference in finite samples, since larger values of T_0 are associated with lower variability of agents' beliefs at the beginning of the observed sample. It is possible to account for that in empirical work by using a two-parameter family of DGLS algorithms indexed by both θ and T_0.

confidence intervals match this theoretical prediction rather closely. In the smaller seasonally adjusted data set (1947-2015), the confidence intervals are quite a lot wider, $[0.08, 0.48]$ for IH and $[22, 53]$ for EE, but the conclusions are qualitatively similar. The main takeaway is that confidence intervals are nonempty in all cases, indicating that there are several values of θ for which our simple learning models can match the low frequency variation in y_t endogenously.

As a final illustration, Figure 2 plots alongside y_t the corresponding forecasts y_{t+1}^e evaluated at the lower bounds of the confidence intervals for θ in each specification. We notice that the series $y_{t+1}^{e\ IH} = (1 - \beta)^{-1} x_{t+1}^{e\ IH}(\theta)$ and $y_{t+1}^{e\ EE}$ are very close to each other and seem to capture well the low frequency movement in inflation.

6   Conclusion

We studied the implications of learning in models where endogenous variables depend on agents' expectations. In a prototypical representative-agent forward-looking model with linear learning algorithms, we found that decreasing gain least squares learning can generate strong persistence that is akin to long memory. The degree of long memory (defined in terms of the order of magnitude of the long-run variance of a process) generated by learning is similar to that of a fractionally integrated process of order $d \in (0, 1/2)$. However, the law of motion that we obtain here is non-stationary, and therefore different from that of a stationary fractionally integrated process. The degree of persistence induced by learning depends negatively on the weight agents place on

21

past observations when they update their beliefs. It also depends positively on the magnitude of the feedback from expectations to the endogenous variable, except in the case of infinite-horizon learning, where agents learn only about the law of motion of the exogenous variable. Importantly, long memory arises endogenously without the need for any persistence in the exogenous shocks. This is distinctly different from the behavior of the model under rational expectations, where the memory of the endogenous variable is determined exogenously and the feedback on expectations has no impact. Moreover, our results are obtained without any of the features that have been previously shown in the literature to be associated with long memory, such as structural change, heterogeneity and nonlinearities.

Appendix

A Proofs

Proof of Proposition 1. Equation (10) implies (9). Thus, we look for a solution of
the form

$$y_t = \nu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{12}$$

that satisfies $y_t = \beta E_t(y_{t+1}) + x_t$. Substituting for y_t, $E_t(y_{t+1})$ and x_t in (1) using (12),
the law of iterated expectations and the definition of x_t in Assumption A' yields

$$(1 - \beta)\nu + \sum_{j=0}^{\infty} (\psi_j - \beta\psi_{j+1})\,\epsilon_{t-j} = \mu + \sum_{j=0}^{\infty} \vartheta_j \epsilon_{t-j}.$$

Identifying the coefficients, it follows that $\nu = \mu(1-\beta)^{-1}$ and $\psi_j - \beta\psi_{j+1} = \vartheta_j$ for all
$j \geq 0$, so $\psi_j = \vartheta_j + \beta\psi_{j+1}$. Assumption A' implies that $\sum_{k=j}^{\infty} \beta^{k-j}\vartheta_k$ converges, so the
forward solution $\psi_j = \sum_{k=j}^{\infty} \beta^{k-j}\vartheta_k$ exists. Moreover, $\sum_{j=0}^{\infty} |\psi_j|$ is finite since

$$\sum_{j=0}^{\infty} |\psi_j| \leq \sum_{j=0}^{\infty}\sum_{k=j}^{\infty} |\beta^{k-j}|\,|\vartheta_k| \leq \sum_{j=0}^{\infty} |\beta^j| \sum_{k=0}^{\infty} |\vartheta_k| < \infty.$$

Hence y_t is covariance stationary (see Brockwell and Davis, 2002, Proposition 2.2.1),
and (12) is the unique stationary solution to (9), see Gourieroux *et al.* (1982, sec. 3.2).
Now if $\sum_{j=0}^{\infty} \psi_j \neq 0$ then by Proposition 3.2.1(i) of Giraitis, Koul and Surgailis (2012),
y_t is short memory and $\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) \asymp 1$. If $\sum_{j=0}^{\infty} \psi_j = 0$, then Proposition
3.2.1(iii) of Giraitis et al. (2012) imply that $\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) = o(1)$. ∎

The proofs of Theorems 2 and 3 rely on the following Lemma.[15] We use the con-
vention that for any t and any sequence f_j, $\prod_{j=t+1}^{t} f_j \equiv 1$.

---

[15]We thank an anonymous referee for helping us simplify the presentation of the results.

Lemma 4 *Consider the process z_t given by*

$$z_t = \prod_{j=1}^{t} (1 - \lambda g_j) z_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} (1 - \lambda g_j) g_i x_i, \qquad (13)$$

where $E\left(z_0^2\right) < \infty$, $g_t = \theta/t + O\left(t^{-1-\varsigma}\right)$, $\varsigma > 0$, $\theta > 0$, $\lambda > 0$, and x_t is i.i.d., with $E\left(x_t^2\right) < \infty$. Then,

$$\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} z_t\right) \asymp \begin{cases} T^{\frac{1}{2} - \lambda\theta}, & \text{if } \lambda\theta < \frac{1}{2}, \\[2mm] \sqrt{\log T}, & \text{if } \lambda\theta = \frac{1}{2}, \\[2mm] 1, & \text{if } \lambda\theta > \frac{1}{2}. \end{cases} \qquad (14)$$

Proof. Assumption $g_t = \theta t^{-1} + O\left(t^{-1-\varsigma}\right)$, $\varsigma > 0$, implies there exist $t_1 < \infty$ such that for all $t > t_1$, $\lambda g_t < 1$. Hence for $T > t_1$, letting $S_T = \sum_{t=1}^{T} z_t$,

$$S_T = S_{t_1} + h_{T+1} z_0 \prod_{j=1}^{t_1} (1 - \lambda g_j) + \sum_{t=t_1+1}^{T} \phi_{T,t} x_t$$

where $k_t = \prod_{j=t_1+1}^{t} (1 - \lambda g_j)$, $h_t = \sum_{j=t_1+1}^{t-1} k_j$ and $\phi_{T,t} = g_t \sum_{i=t}^{T} \prod_{j=t+1}^{i} (1 - \lambda g_j) = \frac{g_t}{k_t} (h_{T+1} - h_t)$.

The variance of S_T is given by

$$\mathsf{Var}\left[S_T\right] = \mathsf{Var}\left[S_{t_1}\right] + \mathsf{Var}\left[z_0\right] \left[h_{T+1}^2 \prod_{j=1}^{t_1} (1 - \lambda g_j)^2 + h_{T+1} \psi_{t_1} \right] + \sigma_x^2 \sum_{t=t_1+1}^{T} \phi_{T,t}^2, \qquad (15)$$

where $\sigma_x^2 = \mathsf{Var}\left[x_t\right]$ and $\psi_{t_1} = 2 \prod_{j=1}^{t_1} (1 - \lambda g_j) \sum_{i=1}^{t_1} \prod_{\ell=1}^{i} (1 - \lambda g_\ell)$. By the assumption $E\left[z_0^2\right] < \infty$ and the fact that $t_1 < \infty$, $\mathsf{Var}\left[S_{t_1}\right] = O\left(1\right)$, $\psi_{t_1} = O\left(1\right)$ and $\prod_{j=1}^{t_1} (1 - \lambda g_j)^2 = O\left(1\right)$. Hence, we study next the magnitudes of h_{T+1}, h_{T+1}^2 and $\sum_{t=t_1+1}^{T} \phi_{T,t}^2$ which appear on the right hand side of Expression (15). We show that $\mathsf{Var}\left[S_T\right] \asymp \sum_{t=t_1+1}^{T} \phi_{T,t}^2$.

24

In the following we use extensively the property that as $t \to \infty$, $\sum_{j=1}^{t} j^{-a} \asymp t^{1-a}$

for $a < 1$, $\sum_{j=1}^{t} j^{-1} \asymp \log t$ and $\sum_{j=1}^{t} j^{-a} \asymp 1$ for $a > 1$.

The asymptotic rates of h_t and k_t depend on the value of $\lambda\theta$. For all $j > t_1$, we can

write $\log (1 - \lambda g_j) = -\lambda g_j + O\left(g_j^2\right)$. Moreover, $\sum_{j=t_1+1}^{t} g_j = \theta \log t + O(1)$ as $t \to \infty$.

Hence, $k_t \asymp \exp\left(-\lambda \sum_{j=t_1+1}^{t} g_j + O\left(\sum_{j=t_1+1}^{t} g_j^2\right)\right) \asymp \exp\left(-\lambda\theta \log t + O(1)\right) \asymp t^{-\lambda\theta}$.

Therefore $g_t/k_t \asymp t^{\lambda\theta-1}$ and $h_t = \sum_{j=t_1+1}^{t-1} k_j \asymp \sum_{j=t_1+1}^{t-1} j^{-\lambda\theta}$, which implies

(a1) $h_t \asymp t^{1-\lambda\theta}$ for $\lambda\theta < 1$;

(a2) $h_t \asymp \log t$ for $\lambda\theta = 1$;

(a3) $h_t \asymp 1$ for $\lambda\theta > 1$.

Hence $\phi_{T,t} = \frac{g_t}{k_t}\left(h_{T+1} - h_t\right) \asymp t^{\lambda\theta-1} \sum_{j=t}^{T} j^{-\lambda\theta}$, and

(b1) $\phi_{T,t} \asymp \frac{T^{1-\lambda\theta}-(t-1)^{1-\lambda\theta}}{t^{1-\lambda\theta}}$ for $\lambda\theta < 1$;

(b2) $\phi_{T,t} \asymp \log \frac{T}{t}$ for $\lambda\theta = 1$;

(b3) $\phi_{T,t} \asymp 1$ for $\lambda\theta > 1$, since $t^{\lambda\theta-1} \sum_{j=t}^{T} j^{-\lambda\theta} \leq t^{\lambda\theta-1} \sum_{j=t}^{\infty} j^{-\lambda\theta} < \infty$ because $\sum_{j=t}^{\infty} j^{-\lambda\theta} \sim$

$t^{1-\lambda\theta}/(\lambda\theta - 1)$.

We now turn to $\sum_{t=1}^{T} \phi_{T,t}^2$. First, if $\lambda\theta < 1$, from (b1) we get

$$\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T^{2(1-\lambda\theta)} \sum_{t=t_1+1}^{T} t^{-2(1-\lambda\theta)} - 2T^{1-\lambda\theta} \sum_{t=t_1+1}^{T} \frac{\left(1-\frac{1}{t}\right)^{1-\lambda\theta}}{t^{1-\lambda\theta}} + \sum_{t=t_1+1}^{T} \left(1-\frac{1}{t}\right)^{2(1-\lambda\theta)}$$

$$\asymp \left(T^{2(1-\lambda\theta)} \sum_{t=t_1+1}^{T} t^{-2(1-\lambda\theta)}\right) + T$$

because $\sum_{t=t_1+1}^{T} \frac{(1-1/t)^{1-\lambda\theta}}{t^{1-\lambda\theta}} \leq \sum_{t=t_1+1}^{T} t^{\lambda\theta-1} \asymp T^{\lambda\theta}$ and $\sum_{t=t_1+1}^{T} (1-1/t)^{2(1-\lambda\theta)} \asymp T$.

Hence, we have the following cases:

(c1) if $\lambda\theta < 1/2$, $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T^{2(1-\lambda\theta)} \sum_{t=t_1+1}^{T} t^{-2(1-\lambda\theta)} \asymp T^{2(1-\lambda\theta)}$;

(c2) if $\lambda\theta = 1/2$, $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T \sum_{t=t_1+1}^{T} t^{-1} \asymp T \log T$;

25

$(c3)$ if $\lambda\theta \in (1/2, 1)$, $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T$, since $\sum_{t=t_1+1}^{T} t^{-2(1-\lambda\theta)} \asymp T^{1-2(1-\lambda\theta)}$.

Next, if $\lambda\theta = 1$, then by $(b2)$ we have $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp \sum_{t=t_1+1}^{T} (\log T/t)^2 \asymp \sum_{t=t_1+1}^{T} \log^2 T$

$-2\log T \log t + \log^2 t$, where $\sum_{t=t_1+1}^{T} \log t \sim T(\log T - 1)$ and $\sum_{t=t_1+1}^{T} \log^2 t \sim T(\log^2 T - 2\log T + 2)$.

Combining all terms yields

$(c4)$ $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T$ for $\lambda\theta = 1$. Finally, by $(b3)$ we obtain

$(c5)$ $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T$ for $\lambda\theta > 1$.

Hence, $(c3)$–$(c5)$ show that $\sum_{t=t_1+1}^{T} \phi_{T,t}^2 \asymp T$ for all values $\lambda\theta > 1/2$.

Comparing $(a1)$-$(a3)$ to $(c1)$-$(c5)$, we notice that $h_T \asymp h_{T+1}^2$ and $h_{T+1}^2 \asymp \sum_{t=t_1+1}^{T} \phi_{T,t}^2$.

Expression (14) and the Lemma follow from the rates $(c1)$-$(c5)$. $\blacksquare$

Proof of Theorem 2. Consider the partial sum of y_t, $S_T = \sum_{t=1}^{T} y_t = \sum_{t=1}^{T} (\beta a_t + x_t)$.

Using expressions (1) and (6), $a_t = \frac{1-g_t}{1-\beta g_t} a_{t-1} + \frac{g_t}{1-\beta g_t} x_t$ or

$$a_t = \prod_{j=1}^{t} \left(1 - \frac{(1-\beta) g_j}{1 - \beta g_j}\right) a_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \frac{(1-\beta) g_j}{1 - \beta g_j}\right) \frac{g_i x_i}{1 - \beta g_i}. \qquad (16)$$

When $g_i \to 0$, $\frac{g_i}{1-\beta g_i} = g_i + o(g_i)$, so the order of magnitude of a_t is the same as that of

$$a_t^* = \prod_{j=1}^{t} (1 - (1-\beta) g_j) a_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} (1 - (1-\beta) g_j) g_i x_i. \qquad (17)$$

Hence, we can infer the order of magnitude of $\mathsf{Var}(S_T)$ from that of $\mathsf{Var}(S_T^*)$, where

$S_T^* = \sum_{t=1}^{T} (\beta a_t^* + x_t)$. Now,

$$\mathsf{Var}(S_T^*) = \beta^2 \mathsf{Var}\left(\sum_{t=1}^{T} a_t^*\right) + \mathsf{Var}\left(\sum_{t=1}^{T} x_t\right) + 2\beta \mathsf{Cov}\left(\sum_{t=1}^{T} a_t^*, \sum_{t=1}^{T} x_t\right).$$

The order of magnitude of the first term on the right-hand side is obtained from Lemma

4 with $z_t = a_t^*$ and $\lambda = 1-\beta$. The second term is $O(T)$ by Assumption A, and therefore,

its order of magnitude is weakly dominated by that of the first term. The third term is

dominated by the first two terms by the Cauchy-Schwarz inequality. Hence, the result of the theorem follows from expression (13) with $\lambda = 1 - \beta$. ∎

Proof of Theorem 3. Consider the partial sum of y_t, $S_T = \sum_{t=1}^{T} y_t$. Expressions (2) and (7) imply that

$$y_t = x_t + \frac{\beta}{1 - \beta} b_t, \quad \text{with } b_t = \prod_{j=1}^{t} (1 - g_j) b_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} (1 - g_j) g_i x_i. \qquad (18)$$

Hence,

$$\mathsf{Var}\,(S_T) = \frac{\beta^2}{(1 - \beta)^2} \mathsf{Var}\left(\sum_{t=1}^{T} b_t\right) + \mathsf{Var}\left(\sum_{t=1}^{T} x_t\right) + \frac{2\beta}{1 - \beta} \mathsf{Cov}\left(\sum_{t=1}^{T} b_t, \sum_{t=1}^{T} x_t\right).$$

Theorem 3 follows from Lemma 4 with $\lambda = 1$. The order of magnitude of the first term on the right-hand side is obtained from Lemma 4 with $z_t = b_t$ and $\lambda = 1$. The second term is $O\,(T)$ by Assumption A, and therefore, its order of magnitude is weakly dominated by that of the first term. The third term is dominated by the first two terms by the Cauchy-Schwarz inequality. ∎

B Extension of Theorems 2 and 3 under Assumption A'

By replacing Lemma 4 in the proofs of Theorems 2 and 3 with Lemma 4' below we establish that the results of the theorems hold more generally under Assumption A'.

Lemma 4' *Consider the process z_t given by (13) where $E\,(z_0^2) < \infty$, $g_t = \theta/t + O\left(t^{-1-\varsigma}\right)$, $\varsigma > 0$, $\theta > 0$, $\lambda > 0$, and x_t satisfies Assumption A'. Then, eq. (14) holds.*

Proof. For notational simplicity, we assume here that t_1 defined in the proof of Lemma 4 is equal to zero, since $t_1 < \infty$ does not impact asymptotic orders of magnitude. Compared to the proof of Lemma 4, $\mathsf{Var}\,(S_T)$ contains the term $2\sum_{t=1}^{T-1}\phi_{T,t}\sum_{i=1}^{T-t}\phi_{T,t+i}\gamma_x\,(i)$ in addition to those in eq. (15).

First, note that $\sum_{j=0}^{\infty}|\vartheta_j| < \infty$ in Assumption A' implies $\sum_{j=0}^{\infty}|\gamma_x\,(j)| < \infty$ (see, e.g., Hamilton, 1994, p. 70). It suffices to establish that

$$\sum_{j=1}^{T-t}\phi_{T,t+j}\gamma_x\,(j) = O\,(\phi_{T,t})\,. \tag{19}$$

If $\lambda\theta < 1$, then by $(b1)$ in the proof of Lemma 4, $\phi_{T,t} \asymp (T/t)^{1-\lambda\theta} - ((t-1)\,/t)^{1-\lambda\theta}$. Hence, there exists $C > 0$ such that

$$\left|\sum_{j=1}^{T-t}\phi_{T,t+j}\gamma_x\,(j)\right| \le C\sum_{j=1}^{T-t}\left[\left(\frac{T}{t+j}\right)^{1-\lambda\theta} - \left(\frac{t+j-1}{t+j}\right)^{1-\lambda\theta}\right]|\gamma_x\,(j)|$$

$$\le C\left(\frac{T}{t}\right)^{1-\lambda\theta}\sum_{j=1}^{\infty}|\gamma_x\,(j)| = O\left(\left(\frac{T}{t}\right)^{1-\lambda\theta}\right)$$

which proves (19).

If $\lambda\theta = 1$, then by $(b2)$ in the proof of Lemma 4, $\phi_{T,t} \asymp \log\frac{T}{t}$ and there exists $C > 0$ such that $\left|\sum_{j=1}^{T-t}\phi_{T,t+j}\gamma_x\,(j)\right| \le C\log\frac{T}{t}\sum_{j=1}^{\infty}|\gamma_x\,(j)| = O\left(\log\frac{T}{t}\right)$, which proves (19).

Finally, if $\lambda\theta > 1$, then by $(b3)$ in the proof of Lemma 4, $\phi_{T,t} \asymp 1$, so $\sum_{j=1}^{T-t}\phi_{T,t+j}\gamma_x\,(j) = O\left(\sum_{j=1}^{T-t}|\gamma_x\,(j)|\right) = O\,(1)\,,$ which proves (19). ∎

References

Abadir, K. M. and G. Talmain (2002). Aggregation, persistence and volatility in a macro model. *Review of Economic Studies 69*(4), 749–79.

Adam, K., A. Marcet, and J. P. Nicolini (2016). Stock market volatility and learning. *Journal of Finance 71*, 33–82.

Andrews, D. W. K. and D. Pollard (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review / Revue Internationale de Statistique 62*(1), pp. 119–132.

Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman & Hall.

Blanchard, O. J. and S. Fischer (1989). *Lectures on Macroeconomics*. MIT press.

Brockwell, P. J. and R. A. Davis (2002). *Introduction to Time Series and Forecasting*. Springer.

Campbell, J. Y. and R. J. Shiller (1987). Cointegration and tests of present value models. *Journal of Political Economy 95*, 1062–1088.

Chambers, M. J. (1998). Long memory and aggregation in macroeconomic time series. *International Economic Review 39*(4), pp. 1053–1072.

Chevillon, G., A. Hecq, and S. Laurent (2015). Long memory through marginalization of large systems and hidden cross-section dependence. Working paper, ESSEC Business School.

Chevillon, G. and S. Mavroeidis (2017). Perpetual learning, apparent long memory and self-confirming equilibria. mimeo.

Clarida, R., J. Galí, and M. Gertler (1999). The science of monetary policy: A New Keynesian perspective. *Journal of Economic Literature 37*, 1661–1707.

Davidson, J. and P. Sibbertsen (2005). Generating schemes for long memory processes: regimes, aggregation and linearity. *Journal of Econometrics 128*(2), 253–82.

Davidson, J. and T. Teräsvirta (2002). Long memory and nonlinear time series. *Journal of Econometrics 110*(2), 105–12.

Diebold, F. X. and A. Inoue (2001). Long memory and regime switching. *Journal of Econometrics 105*(1), 131–159.

Eusepi, S. and B. Preston (2011). Expectations, learning, and business cycle fluctuations. *American Economic Review 101*(6), 2844–72.

Evans, G. W. and S. Honkapohja (2001). *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Evans, G. W. and S. Honkapohja (2009). Expectations, learning and monetary policy: An overview of recent rersearch. In K. Schmidt-Hebbel and C. Walsh (Eds.), *Monetary Policy under Uncertainty and Learning*, pp. 27–76. Santiago: Central Bank of Chile.

Galí, J. and M. Gertler (1999). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics 44*, 195–222.

Geweke, J. and S. Porter-Hudak (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis 4*, 221–38.

Giraitis, L., H. L. Koul, and D. Surgailis (2012). *Large sample inference for long memory processes*, Volume 201. World Scientific.

Gourieroux, C., J. J. Laffont, and A. Monfort (1982). Rational expectations in dynamic linear models: Analysis of the solutions. *Econometrica 50*(2), pp. 409–425.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227–238.

Granger, C. W. J. and Z. Ding (1996). Varieties of long memory models. *Journal of econometrics 73*(1), 61–77.

Haldrup, N. and R. Kruse (2014). Discriminating between fractional integration and spurious long memory. Technical report, University of Aarhus.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Malmendier, U. and S. Nagel (2016). Learning from inflation experiences. *Quarterly Journal of Economics 131*, 53–87.

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

Marcet, A. and T. J. Sargent (1995). Speed of convergence of recursive least squares: Learning with autoregressive moving-average perceptions. In A. Kirman and M. Salmon (Eds.), *Learning and Rationality in Economics*, pp. 179–215. Oxford: Basil Blackwell.

McCallum, B. T. (1983). On non-uniqueness in rational expectations models: An attempt at perspective. *Journal of Monetary Economics 11*, 139–168.

Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics 54*(7), 2065–2082.

Miller, J. I. and J. Y. Park (2010). Nonlinearity, nonstationarity, and thick tails: How they interact to generate persistence in memory. *Journal of Econometrics 155*(1), 83–89.

Müller, U. and M. W. Watson (2008). Testing models of low-frequency variability. *Econometrica 76*(5), 979–1016.

Parke, W. R. (1999). What is fractional integration? *Review of Economics and Statistics 81*(4), 632–638.

Perron, P. and Z. Qu (2007). An analytical evaluation of the log-periodogram estimate in the presence of level shifts. working paper, Boston University.

Preston, B. (2005). Learning about monetary policy rules when long-horizon expectations matter. *International Journal of Central Banking 1*(2), 81–126.

Preston, B. (2006). Adaptive learning, forecast-based instrument rules and monetary policy. *Journal of Monetary Economics 53*(3), 507–535.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics 23*, 1630–61.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. 42*(1), 43–47.

Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics*. Oxford University Press.

Sargent, T. J. (1999). *The conquest of American inflation*. Princeton University Press.

Schennach, S. (2013). Long memory via networking. Working paper cwp13/13, CEMMAP.

Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory 26*, 501–540.

Shimotsu, K. and P. C. B. Phillips (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics 33*(4), 1890–1933.

White, H. (2000). *Asymptotic Theory for Econometricians*. Academic Press Inc. 2nd ed.

Zaffaroni, P. (2004). Contemporaneous aggregation of linear dynamic models in large economies. *Journal of Econometrics 120*(1), 75 – 102.
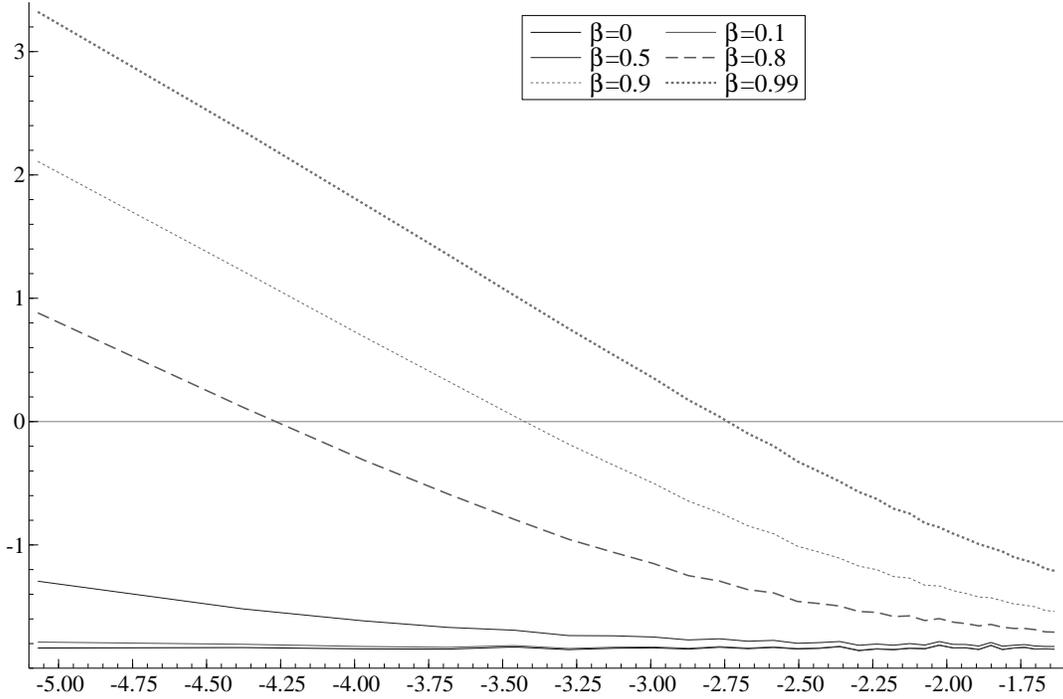
Figure 1: Monte Carlo averages of the log periodogram of $y_t = \beta y_{t+1}^e + x_t$, where x_t is iid N(0,1) and y_{t+1}^e is given by RLS, against the log of the first $\sqrt{T}$ Fourier frequencies, with $T = 1,000$.

β	Mean of $\widehat{d}$		Pr(Reject $d = 0$)		Pr(Reject $d = 1$)	
	GPH	Whittle	GPH	Whittle	GPH	Whittle
0.00	0.001	-0.011	0.075	0.069	0.938	0.996
0.10	0.006	-0.007	0.081	0.077	0.924	0.993
0.50	0.055	0.039	0.179	0.182	0.797	0.951
0.80	0.291	0.245	0.656	0.677	0.563	0.755
0.90	0.438	0.378	0.805	0.817	0.467	0.635
0.99	0.573	0.510	0.890	0.899	0.376	0.520

Table 1: Estimates and rejection frequencies of tests on the long memory parameter d for $y_t = \beta y_{t+1}^e + x_t$, under RLS learning.

	NSA 1913(2)-2015(8)	SA 1947(1)-2015(8)
$\hat{d}$ of inflation	0.38	0.42
st.error	0.085	0.095
90% CI for θ in EE	[8,27]	[22,53]
90% CI for θ in IH	[0.03,0.08]	[0.08,0.48]

Table 2: Estimates of memory parameter of U.S. monthly CPI inflation and confidence intervals for the forgetting factor θ in the learning algorithm. $\widehat{d}$ is 2EWL of Shimotsu and Phillips (2005). CIs are based on the 2EWL estimator.
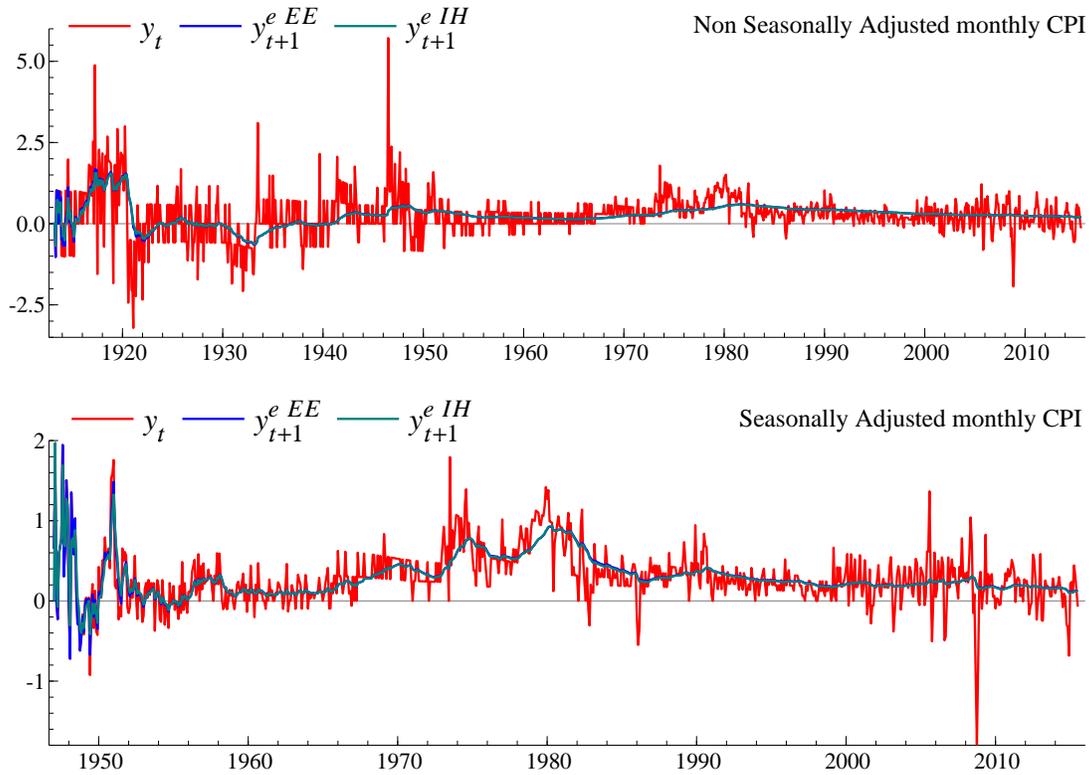

Figure 2: U.S. monthly CPI inflation y_t along its expectations under EE and IH learning, computed for the smallest values of the forgetting factor θ reported in Table 2.