Learning can generate Long Memory*

Guillaume Chevillon† Sophocles Mavroeidis‡
ESSEC Business School University of Oxford

October 10, 2013

Abstract

We consider a prototypical representative-agent forward-looking model, and study the low frequency variability of the data when the agent's beliefs about the model are updated through linear learning algorithms. We find that learning in this context can generate strong persistence. The degree of persistence depends 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. Under recursive least squares learning, long memory arises when the coefficient on expectations is sufficiently large. Under discounted least squares learning, long memory provides a very good approximation to the low-frequency variability of the data. Hence long memory arises endogenously, due to the self-referential nature of the model, without any persistence in the exogenous shocks. This is distinctly different from the case of rational expectations, where the memory of the endogenous variable is determined exogenously. Finally, this property of learning is used to shed light on some well-known empirical puzzles.

JEL Codes: C1, E3;

Keywords: Learning, Long Memory, Persistence, Present-Value Models.

*We would like to thank Karim Abadir, Richard Baillie, Jess Benhabib, George Evans, Seppo Honkapohja, Peter Howitt, Cliff Hurvich, Rustam Ibragimov, Frank Kleibergen, Guy Laroque, Ulrich Müller, Mark Watson, Ken West, as well as the participants in the 2010 NBER summer institute for helpful comments and discussions. We also benefited from comments received at the Nordic Econometric Meeting in Lund, the Netherlands Econometric Study Group in Leuven, the ESEM in Oslo, EC^2 in Toulouse, the NBER/NSF Time Series Conference at Texas A&M University, as well as from seminar participants at Cambridge, CREST, Durham, ESSEC, the Federal Reserve Bank of New York, GREQAM, NYU, Nottingham, Oxford and Rotterdam. Mavroeidis would like to thank the European Commission for research support under a FP7 Marie Curie Fellowship CIG 293675. Chevillon acknowledges research support from CREST.

†ESSEC Business School, Department of Information Systems, Decision Sciences and Statistics, Avenue Bernard Hirsch, BP50105, 95021 Cergy-Pontoise cedex, France. Email: chevillon@essec.edu.

‡Department of Economics and INET at Oxford, University of Oxford, Manor Road, Oxford, OX1 3UQ, United Kingdom. Email: sophocles.mavroeidis@gmail.com.

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 the new Keynesian policy model, 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). In a related paper, Chevillon, Massmann and Mavroeidis (2010) showed that the persistence induced by learning can be so strong as to invalidate conventional econometric methods of estimation and inference.

The objective of this paper is to point out the connection between learning and long memory. In particular, we show that in certain economic models, replacing rational expectations with certain types of learning can generate long memory. 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 heavily 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 consider two applications on excess return predictability (see Stambaugh, 1999) and the forward premium anomaly (see Engel, 1996), where we find that learning can provide an endogenous explanation for the observed long memory of the dividend–price ratio and various forward premia.

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); and (iii) some form of nonlinearity (see, *e.g.,* Davidson and Sibbertsen, 2005, and Miller and Park, 2010). 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.

The paper is organized as follows. Section 2 presents 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 discuss two empirical applications. 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 $\lim_{x\to a} f(x)/g(x) = 1$; $O(\cdot)$ and $o(\cdot)$ denote standard orders of magnitude; and $f(x) = \mathcal{O}(g(x))$ means "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

Consider the following forward-looking model that links an endogenous variable y_t to an exogenous process x_t:

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

where y^e_{t+1} denotes the expectation of y_{t+1} conditional on information up to time t. 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 dependence discussed below.

Under rational expectations, $y^e_{t+1} = E_t(y_{t+1})$, where E_t denotes expectations based on the true law of motion of y_t. It is well-known that when $|\beta| < 1$ and $\lim_{T\to\infty} E_t(y_T) < \infty$, the rational expectations equilibrium (REE) satisfies

$$y_t = \sum_{j=0}^{\infty} \beta^j E_t(x_{t+j}), \tag{2}$$

provided this sum converges, which depends on the properties of x_t. Under adaptive learning (Evans and Honkapohja, 2001, 2009), agents form expectations based on some perceived law of motion (PLM) for the process y_t, whose parameters are recursively estimated using information available to them. The simplest PLM is the mean-plus-noise model

$$y_t = \alpha + \epsilon_t, \tag{3}$$

where α is an unknown parameter, and ϵ_t is an identically and indepently distributed (*i.i.d.*) shock.[1] 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^e_{t+1} is given by a recursive estimate of α. The classic learning algorithm is recursive least squares (RLS): $y^e_{t+1} = \frac{1}{t}\sum_{i=1}^{t} y_i$. This is a member of the class of weighted least squares algorithms that are defined as the solution to the minimization problem

$$y^e_{t+1} = \underset{a}{\operatorname{argmin}} \sum_{j=0}^{t-1} w_{t,j}\left(y_{t-j} - a\right)^2, \quad \sum_{j=0}^{t-1} w_{t,j} = 1. \tag{4}$$

RLS corresponds to $w_{t,j} = t^{-1}$. 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 $\bar{g} \in (0,1)$.

An alternative characterization of learning in the literature is based on stochastic recursive algorithms (see Evans and Honkapohja, 2001, chapter 6). Consider a slight generalization of the PLM (3) to allow for *perceived* shifts in the mean:

$$y_t = \alpha_t + \epsilon_t, \tag{5a}$$

$$\alpha_t = \alpha_{t-1} + v_t, \quad t \geq 1, \tag{5b}$$

[1] This PLM nests the rational expectations equilibrium that arises when $E_t(x_{t+j})$ is constant for all t, j. Otherwise, it can be interpreted as a restricted perceptions equilibrium (RPE), see Sargent (1993).

where $\alpha_0 = \alpha$; ϵ_t and v_t are *i.i.d.* with mean zero and finite variances, and define the signal-to-noise ratio $\tau_t = \mathsf{Var}(v_t)/\mathsf{Var}(\epsilon_t)$. Under the PLM (5), y^e_{t+1} is given by a function of current and past values of y_t that estimates α_t. If the errors ϵ_t, v_t are Gaussian, the optimal estimate of α_t, denoted by a_t, is given by the Gaussian Kalman Filter (see Durbin and Koopman, 2008):

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

$$g_t = \frac{g_{t-1} + \tau_t}{1 + g_{t-1} + \tau_t}, \quad t \geq 2, \;\; g_1 = \frac{\sigma_0^2 + \tau_1}{1 + \sigma_0^2 + \tau_1} \tag{6b}$$

with a_0 and σ_0^2 measuring the mean and variance of agents' prior beliefs about α. The parameter σ_0^2 can also be interpreted as inversely related to agents' confidence in their prior expectation of α, given by a_0. g_t is the so-called gain sequence. When $g_t = \bar{g}$ for all t, the algorithm is called constant gain least squares (CGLS). RLS arises as a special case when $\tau_t = 0$ for all t and $\sigma_0^2 \to \infty$, so that $g_t = 1/t$. This 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 $\nu \in (0, 1]$, as discussed Evans and Honkapohja (2001, chapter 7). Malmendier and Nagel (2013) 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 A in the Appendix for details.

The above learning algorithms can be expressed as linear functions of past values of y_t with possibly time-varying coefficients:

$$y^e_{t+1} = \sum_{j=0}^{t-1} \kappa_{t,j} y_{t-j} + \varphi_t. \tag{7}$$

where the term φ_t represents the impact of the initial beliefs. Our main motivation for focusing our attention on linear learning algorithms is 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). We use a representative agent framework to avoid inducing long memory through heterogeneity and aggregation, as in, *e.g.*, Granger (1980), Abadir and Talmain (2002), Zaffaroni (2004) and Schennach (2013).

We define the polynomial $\kappa_t(L) = \sum_{j=0}^{t-1} \kappa_{t,j} L^j$ where L is the lag operator. To quantify how much agents discount past observations when forming expectations, we use the mean lag of κ_t, which is defined as

$$m(\kappa_t) = \frac{1}{\kappa_t(1)} \sum_{j=1}^{t-1} j\kappa_{t,j}. \tag{8}$$

The magnitude of $m(\kappa_t)$ relative to the sample size can be used to measure the 'length' of the learning window. We show below that this drives the memory of the process that is induced by learning dynamics. The following definition provides our measure of the length of the learning window.

Definition LW (length of learning window) *Suppose there exist scalars $m_\kappa > 0$ and $\delta_\kappa \geq 0$ such that $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa}$, as $t \to \infty$. Then, δ_κ is referred to as the length of the learning window. The learning window is said to be short when $\delta_\kappa = 0$ and long otherwise.*

In the paper, we make the following assumptions about the general linear learning algorithm (7):

Assumption A.

A.1. κ_t is nonstochastic;

A.2. $\{\kappa_{t,j}\}$ is absolutely summable with $\kappa_t(1) \leq 1$ for all t;

A.3. There exists $m_\kappa > 0$ and $\delta_\kappa \in [0,1]$ such that $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa}$, as $t \to \infty$.

Assumption A.1 could be relaxed to allow $\{\kappa_{t,j}\}$ to be stochastic, provided that it is independent of $\{x_t\}$, in which case our results would be conditional on almost all realizations of $\{\kappa_{t,j}\}$. It precludes cases in which $\kappa_{t,j}$ depends on lags of y_t, such as when the PLM is an autoregressive model, because in those cases the learning algorithm is nonlinear.[2]

Assumption A.2 is a common feature of most learning algorithms. It implies in particular that $\kappa_{t,t-1} \to 0$ as $t \to \infty$. Under assumption A.3 $\lim_{t\to\infty} \frac{\log m(\kappa_t)}{\log t}$ exists. This precludes cases where there exists a slowly varying function S_κ (i.e., where $\lim_{t\to\infty} S_\kappa(\lambda t)/S_\kappa(t) = 1$ for $\lambda > 0$) such that $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa} S_\kappa(t)$. This is inconsequential to our analysis (although it will exclude some parameter values in Section 3) but simplifies the exposition since $\delta_\kappa = 0$ implies here that $m(\kappa_t)$ is bounded.

[2] Assumption A.1 also avoids the issue of generating fat tails through a random coefficient autoregressive model as in Benhabib and Dave (2013).

Learning Algorithm		$\kappa_{j,t}$	gain	δ_κ
DGLS	$\theta \geq 1$	$\theta \frac{\Gamma(t+1-\theta)}{\Gamma(t-j+1-\theta)} \frac{\Gamma(t-j)}{\Gamma(t+1)}$	$\min\left(\frac{\theta}{t}, 1\right)$	1
RLS	$\theta = 1$	$t^{-1} 1_{\{j<t\}}$	$\frac{1}{t}$	1
CGLS	$\bar{g} \in (0,1)$	$\bar{g}\,(1-\bar{g})^j$	$\bar{g}$	0
CGLS with small gain	$\bar{g}_T = c_g T^{-\lambda}$	$\bar{g}_T\,(1-\bar{g}_T)^j$	$\bar{g}_T$	λ
HWLS	$\lambda < 1$	$j^{\lambda-2}/\zeta\,(2-\lambda)$	-	$\max(0,\lambda)$
	$\lambda \in (1,2)$	$(\lambda-1)\, j^{\lambda-2} t^{1-\lambda}$	-	1

Table 1: Examples of Weighted Least-Squares Learning algorithms, with corresponding coefficients ($\kappa_{t,j}$), gains and learning window lengths (δ_κ). $\zeta(\cdot)$ denotes Riemann's Zeta function and $\Gamma(\cdot)$ is the Gamma function. DGLS: Decreasing Gain Least Squares; RLS: Recursive Least Squares; CGLS: Constant Gain Least Squares; HWLS: Hyperbolically Weighted Least Squares.

We list the learning algorithms we study later in the paper in Table 1, where we also specify the length of the learning window for each algorithm. The first two algorithms are DGLS, and they are analyzed in Section 3.2. Both are long window algorithms as shown in Section A in the Appendix. The next two algorithms are CGLS, discussed in Section 3.3. The last set of algorithms are weighted least squares algorithms with hyperbolically decaying weights, analyzed in Section 3.4.

Next, 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 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.[3] If this does not hold, we will say that the process exhibits long memory.[4]

[3] 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).

[4] This is also the definition adopted by Diebold and Inoue (2001) in their study of the connection between

Analogously to our previous discussion of the length of the learning window, we can also define the 'degree of memory' of a process z_t by the parameter d (when it exists) such that

$$\mathsf{sd}\left(T^{-1/2}S_T\right) = \mathcal{O}\left(T^d\right), \quad \text{where } S_T = \sum_{t=1}^{T} z_t. \tag{9}$$

Definition LM (long memory) *The process z_t exhibits long memory if $d > 0$ in (9).*[5]

The above definition applies generally to any stochastic process that has finite second moments (which we assume in this paper). For a covariance stationary process, where the autocorrelation function (ACF) is a common measure of persistence, short memory requires absolute summability of its autocorrelation function, or a finite spectral density at zero. Thus, long memory arises when the autocorrelation coefficients are non-summable (typically if they decay hyperbolically), or the spectrum has a pole at frequency zero. This gives rise to alternative definitions of d based on the ACF and spectral density that are equivalent to definition LM for covariance stationary processes, see Section H in the Appendix. When relevant, we also provide in Section H results for these different characterizations of long memory.

Finally, we need to make some assumptions about the forcing variable x_t. These are given by Assumption B below.

Assumption B. There exists an *i.i.d.* process ϵ_t with $E\left|\epsilon_t\right|^r < \infty$ for $r > 2$ and such that $x_t = \sum_{j=0}^{\infty} \vartheta_j \epsilon_{t-j}$, with $\sum_{j=0}^{\infty} \vartheta_j \neq 0$ and $\sum_{j=0}^{\infty} j\left|\vartheta_j\right| < \infty$.

Assumption B characterizes a typical covariance stationary process with short memory and is found in Perron and Qu (2007, Assumption 1) and Perron and Qu (2010); it is weaker than Assumptions LP of Phillips (2007) and Magdalinos and Phillips (2009) and constitutes a version of Stock (1994, Assumptions (2.1)-(2.3)) with independent homoskedastic innovations ϵ_t. The assumption ensures x_t satisfies a functional central limit theorem (Phillips and Solo, 1992, theorem 3.4). This assumption includes all covariance stationary processes that admit a finite-order invertible autoregressive moving average (ARMA) representation, and

structural change and long memory.

[5]In the context of nonlinear cointegration, Gonzalo and Pitarakis (2006) have introduced the terminology "summable of order d" for processes that satisfy the definition given in equation (9) above, see also Berenguer-Rico and Gonzalo (2013).

therefore have exponentially decaying autocovariances, but it also includes more persistent short memory processes whose autocovariances decay at slower-than-exponential rates. Assumption B rules out processes with $0 < \left|\sum_{j=0}^{\infty} \vartheta_j\right| < \infty$ and $\left|\sum_{j=0}^{\infty} j\vartheta_j\right| = \infty$, for even though these processes have $d = 0$ in the definition (9), they are difficult to distinguish from long-memory processes in finite samples, as their spectral density is not differentiable at the origin (see Stock, 1994, Sections 2.1 and 2.5).

3 Analytical results

This section provides our main results. We start by showing that long memory cannot arise endogenously under RE. We then analyze the impact of learning on the memory of the resulting process. We start with DGLS learning and then consider the case of CGLS learning in the empirically relevant case where the gain is small. Finally, we look at general learning algorithms whose coefficients are time-invariant, *i.e.*, $\kappa_{t,j} = \kappa_j$ for all t in (7).

3.1 Rational Expectations

The following result shows that long memory cannot arise endogenously under rational expectations when x_t follows a short-memory process described by Assumption B.

Proposition 1 *Suppose x_t satisfies Assumption B, and $y_t = \sum_{j=0}^{\infty} \beta^j E_t x_{t+j}$ with $\beta \in (-1, 1]$. Then,* $\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) = O(1)$.

Note, that this result holds even in the case $\beta = 1$. Hence the magnitude of the feedback that expectations have on the process plays no role for the memory of the process under RE. As we will see below, this is very different from what happens under learning, because under learning the memory of y_t crucially depends on the proximity of β to 1.

3.2 Decreasing gain least squares

When agents learn using DGLS, the learning algorithm has time-varying coefficients and the resulting process y_t is nonstationary. It is well-known that for the class of learning algorithms we consider, decreasing gain learning in this model converges to the REE, see, *e.g.*, Evans and Honkapohja (2001, theorem 7.10). Hence y_t tends to a weakly dependent process. Yet the convergence can be so slow that the autocorrelation of the process y_t decreases very slowly and y_t may exhibit long memory. To gain some intuition for this, consider the impulse response function (IRF) of y_{t+j} with respect to x_t under RLS learning. It is shown in Section C in the Appendix that, if x_t is *i.i.d.* and as $t, j/t$ get large,

$$\frac{\partial y_{t+j}}{\partial x_t} \sim \beta t^{-\beta} j^{-(1-\beta)}. \tag{10}$$

Expression (10) shows that the IRF is time-varying, as expected, and it decays hyperbolically in j. Moreover, the closer β is to unity, the slower the decay of the response for any given t. Expression (10) also shows the persistence is transitory since $\frac{\partial y_{t+j}}{\partial x_t} \to 0$ as $t \to \infty$. Yet when β is sufficiently close to unity, convergence is slow enough for the process y_t to exhibit long memory. The above claim is formally established for the DGLS learning algorithm in the following result.

Theorem 2 *Consider the model $y_t = \beta y^e_{t+1} + x_t$, with $y^e_{t+1} = a_t$ as given in equation (6) where $g_t \sim \theta/t$, $\theta > 0$, $a_0 = O_p(1)$ and x_t satisfies Assumption B. Then, as $T \to \infty$,*

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

The theorem shows that the process exhibits long memory of degree $d \in \left(0, \frac{1}{2}\right]$ when $\beta > 1 - \frac{1}{2\theta}$. The degree of memory is $\max(1/2 - \theta(1-\beta), 0)$. For RLS ($\theta = 1$) this specializes to $\max\left(\beta - \frac{1}{2}, 0\right)$. 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 a constant when $\beta < 1$, asymptotic normality of y^e_{t+1} is only established when $\beta < 1 - \frac{1}{2\theta}$ (Evans and Honkapohja, 2001, theorem 7.10). The long memory that arises when $\beta \geq 1 - \frac{1}{2\theta}$ explains why the standard central limit theorem does not apply to agents' estimators. When $\beta = 1$, learning does not converge and persistence is strongest in that case.

3.3 Constant gain least squares

Another leading example of a learning algorithm that features prominently in the empirical literature is CGLS, or perpetual learning. For fixed gain, CGLS is clearly a short-window algorithm, but this is not an appropriate characterization when the gain parameter is small relative to the sample size. To make this precise, we consider a local-to-zero asymptotic nesting where the gain parameter goes to zero with the sample size.

The CGLS algorithm on the mean-plus-noise PLM (5) makes y_{t+1}^{e} an exponentially weighted moving average of past y_j, $j \leq t$. Specifically, $y_{t+1}^{e} = a_t$, where

$$a_t = \left(\frac{1-\bar{g}}{1-\beta\bar{g}}\right)^t a_0 + \frac{\bar{g}}{1-\beta\bar{g}} \sum_{i=1}^{t} \left(\frac{1-\bar{g}}{1-\beta\bar{g}}\right)^{t-i} x_i. \tag{11}$$

So, if β is close to unity or $\bar{g}$ close to zero such that $(1-\bar{g})/(1-\beta\bar{g}) \approx 1$, a_t exhibits near unit-root behavior (see Bobkoski, 1983, Phillips, 1987). Yet, a small $\bar{g}$ appearing before the summation attenuates the stochastic trend in a_t.

To characterize the dynamics of y_t when β is large and $\bar{g}$ is close to its boundaries, we follow and extend the local-asymptotic approach of Chevillon *et al.* (2010). This constitutes a nesting in which parameters are expressed in relation to the sample size. We let $1-\beta = c_\beta T^{-\nu}$ and $\bar{g} = c_g T^{-\lambda}$ for $(\nu, \lambda) \in [0,1]^2$ and c_β, c_g strictly positive real scalars.[6] Formally, this framework means that the stochastic process of y is a triangular array $\{y_{t,T}\}_{t \leq T}$. However, we shall omit the dependence of β, $\bar{g}$ and y_t on T for notational simplicity.

Section E in the Appendix shows that the mean lag of the learning algorithm satisfies

$$m(\kappa_T) = \mathcal{O}\left(T^{\lambda}\right), \tag{12}$$

so the length of the learning window δ_κ is equal to λ. Hence, $\lambda > 0$, implying $\bar{g} \to 0$, corresponds to long-window learning, while $\lambda = 0$ corresponds to short window learning. The following theorem gives the implications for the memory of y_t.

Theorem 3 *Consider the model $y_t = \beta y_{t+1}^{e} + x_t$, where $y_{t+1}^{e} = a_t$ given by (11), $a_0 = O_p(1)$ and x_t satisfies Assumption B. Suppose that $\beta = 1 - c_\beta T^{-\nu}$ and $\bar{g} = c_g T^{-\lambda}$, where $\nu, \lambda \in [0,1]^2$ and c_β, c_g are positive constants. Then, as $T \to \infty$,*

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) = \mathcal{O}\left(T^{\min(\nu, 1-\lambda)}\right). \tag{13}$$

[6] Chevillon *et al.* (2010) studied only the case where $\nu = \lambda = 1/2$ and x_t is *i.i.d.* They did not consider the implications for the memory of y_t.

Theorem 3 shows that CGLS learning with a large β generates long memory. More specifically, the memory of the process y_t depends on (i) the proximity of β to unity and (ii) the length of the learning window. If $\nu = 0$, *i.e.*, β is 'far' from unity, the process exhibits short memory, irrespective of the length of the learning window. For $\nu > 0$, the memory of the process depends on whether $\nu \leq 1-\lambda$ or $\nu > 1-\lambda$, *i.e.*, on how close β is to unity relative to the length of the learning window. When β is sufficiently close to unity, the memory of the process is determined entirely by the length of the learning window, λ, and is nonincreasing in λ. Persistence is, in fact, strongest when the gain is far from zero, $\lambda = 0$, *i.e.*, when the learning window is short. This may appear counterintutive at first, but it is entirely analogous to what happens in fractionally integrated processes. To gain some intuition, consider the fractional white noise process $(1-L)^d y_t = \varepsilon_t$, where $d \in (-1/2, 1/2)$, $d \neq 0$, and ε_t is white noise. The memory of this process, d, is directly related to the rate of decay of the impulse response function, *i.e.*, the rate of decay of the coefficients of the moving average representation, which is $d-1$.[7] The rate of decay of the autoregressive coefficients is $-d-1$, so it is *inversely* related to d. Therefore, given a unit root in the autoregressive polynomial, a more persistent process is associated with a faster decay of the autoregressive coefficients. In the learning model, this corresponds to a higher discounting of past observations in the learning algorithm, *i.e.*, a shorter learning window.

CGLS learning with a small gain parameter induces behavior that is in some sense close to a rational expectations equilibrium, and it is referred to as 'near-rational expectations' in the literature, see Milani (2007). The smallest gain arises when $\lambda = 1$ in Theorem 3, which leads to short memory. This is exactly what happens under rational expectations, see Proposition 1. So, similarly to rational expectations, learning that is akin to near-rational expectations cannot generate long memory.

Note that CGLS with very small gain is very different from RLS, *i.e.*, the latter is not the limit of the former as the gain parameter goes to zero. Heuristically, near-rational expectations corresponds to the 'limiting' law of motion when RLS learning has converged, and therefore, it misses all the transitional dynamics of RLS, which matter – this is exactly the intuition behind Theorem 2.

[7] See, e.g., Baillie (1996, Table 2).

3.4 Learning algorithms with hyperbolic weights

We can extend the results of the previous section to cover learning algorithms (7) that satisfy Assumption A and have constant coefficients $\kappa_{t,j} = \kappa_j$. CGLS is such an algorithm, but without making the gain parameter local to zero, the weights κ_j decay exponentially and the length of the learning window is short. We now consider situations when weights of the learning algorithm decay hyperbolically in j, so that we can cover long-window algorithms without treating the gain parameter as local to zero. Such algorithms can be motivated as hyperbolically discounted least squares. In some sense, they bridge the gap between RLS (no discounting) and CGLS (exponential discounting). Assumption A.2 implies that $\kappa_j = o\left(j^{-1}\right)$, and the length of the learning window, δ_κ, depends on the rate of decay of the weights. If $\kappa_j = o\left(j^{-2}\right)$, the learning window is short under Assumption A.3, while if $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$, for some $c_\kappa > 0$ and $0 < \delta_\kappa < 1$, the learning window is long, with length δ_κ.[8]

As in the case of CGLS, we use the local asymptotic framework for β, $\beta = 1 - c_\beta T^{-\nu}$, and suppress the triangular array notation for y_t. Unlike CGLS, the weights of the learning algorithm here do not depend on T. Thus, the ensuing results do not cover those of the previous subsection.

For simplicity, we assume that there is an infinite history of $\{y_t\}$ and define the initial beliefs φ_t as $\varphi_t = \sum_{j=t}^{\infty} \kappa_j y_{t-j}$ if $\delta_\kappa \in (1/2, 1)$ and $\Delta\varphi_t = \sum_{j=t}^{\infty} \kappa_j \Delta y_{t-j}$ if $\delta_\kappa \in (0, 1/2)$.[9]

The following result gives the memory properties of the process y_t according to Definition LM.

Theorem 4 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, with $y_{t+1}^e = \kappa(L) y_t$. Suppose x_t satisfies Assumption B and that the learning algorithm $\kappa(\cdot)$ satisfies Assumption A, with $\delta_\kappa \in [0, 1)$, $\delta_\kappa \neq 1/2$, $\kappa(1) = 1$, and $\beta = 1 - c_\beta T^{-\nu}$ with $\nu \in [0, 1]$ and $c_\beta > 0$. Then, as $T \to \infty$,*

$$\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) = \mathcal{O}\left(T^{\min(\nu, 1-\delta_\kappa)}\right).$$

[8] One example of $\kappa(L)$ that satisfies the above assumptions is the operator $L_d = 1 - (1 - L)^d$, $d \in (0, 1)$, such that $\kappa_j \sim c_\kappa j^{-d-1}$, and $\delta_\kappa = 1 - d$, see Granger (1986) and Johansen (2008).

[9] A simplifying assumption often made in the literature is $y_t = 0$ for $t \leq 0$, see, *e.g.*, Diebold and Rudebusch (1991) and Tanaka (1999). Yet, it has been shown that this assumption (which is related to the difference between Type I and Type II Fractional Brownian motions) is not innocuous for the definition of the spectral density, so we avoid it: see Marinucci and Robinson (1999), Davidson and Hashimzade (2008, 2009).

This result is entirely analogous to Theorem 3, where $\delta_\kappa = \lambda$. When β is sufficiently close to unity, $\nu > 1 - \delta_\kappa$, we can derive expressions for the spectral density of y_t at low frequencies and the rate of decay of its autocorrelation function that accord with the alternative common definitions of long memory. These definitions rely either on the hyperbolic behavior of the spectral density in a neighborhood of the origin or on hyperbolic rates of decay of the autocorrelations. The definitions and corresponding theorem are given in Section H in the Appendix. They show that the degree of memory reported in Theorem 4 coincides with the common alternative definitions.

Our results show that the persistence of the process y_t is a function of the relative values of the length of the learning window and the proximity of β to unity. When β is sufficiently close to unity, the memory of the process is determined entirely by the length of the learning window, δ_κ, and is inversely related to the latter. Theorem 4 also shows that if β is well below unity, the memory of y_t is short irrespective of the length of the learning window. So, $\beta \to 1$ is necessary for long memory in y_t under learning algorithms with hyperbolic discounting.

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 and CGLS learning algorithms listed in Table 1. 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 parameters β and $\bar{g}$. 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 estimators of the fractional differencing parameter d.[10] We also report the power of tests of the null hypotheses $d = 0$ and $d = 1$. The number of Monte Carlo replications is 10,000. Additional figures reporting the rate of growth of the variance of partial sums and the densities of estimators of d are available in a supplementary appendix.

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

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

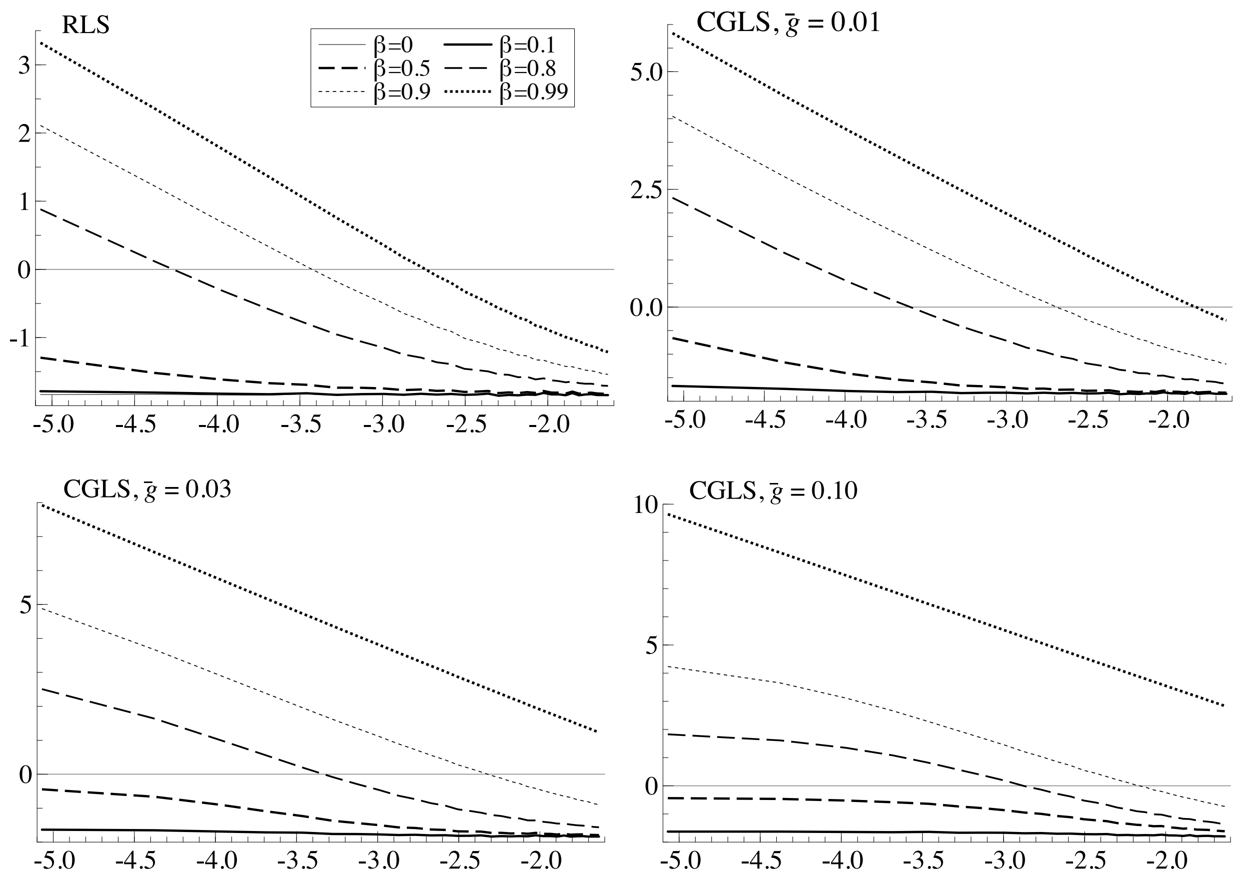

Figure 1: Monte Carlo averages over 10,000 replications of the log periodogram against the log of the first $\sqrt{T}$ Fourier frequencies with $T = 1,000$ observations. The model is $y_t = \beta y^e_{t+1} + x_t$, $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, and y^e_{t+1} is determined by RLS (top left panel) or CGLS (all other panels) learning.

β	Mean of $\widehat{d}$ GPH	Whittle	Pr(Reject $d=0$) GPH	Whittle	Pr(Reject $d=1$) 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 2: The table records estimates and tests on the long memory d for $y_t = \beta y^e_{t+1} + x_t$, under RLS learning. The data is generated as $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, $T = 1000$ and the number of Monte Carlo replications is 10000. GPH is the Geweke & Porter-Hudak (1983) estimator and Whittle is the Robinson (1995) maximum local Whittle likelihood estimator. $\Pr(\text{Reject } d = 0)$ and $\Pr(\text{Reject } d = 1)$ are the empirical rejection 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.

dependence if the log periodogram is linearly decreasing in log ω. When the learning algorithm is RLS, the figure indicates that y_t exhibits long memory for $\beta > 1/2$ and the degree of long memory increases with β. Table 2 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.*, $E(\widehat{d}) \approx \max(0, \beta - 1/2)$. Figure 1 and Table 3 report the corresponding statistics for various values of $\bar{g}$ under CGLS learning. The behavior of $E(\widehat{d})$ as well as $\Pr(\text{Reject } d = 0)$ and $\Pr(\text{Reject } d = 1)$ in terms of β and $\bar{g}$ accords with Theorem 3. Specifically, $E(\widehat{d})$ is increasing in β given $\bar{g}$, and weakly increasing in $\bar{g}$ given β. Since T is fixed, a higher $\bar{g}$ corresponds to a shorter learning window, so the memory of the process is decreasing in the length of the learning window, in accordance with Theorem 3.

Unreported figures (available in the supplementary appendix) 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 both RLS learning and CGLS learning with local parameters. We also present there the densities of the estimators of d which complement the rejection probabilities

$\bar{g}$	β	Mean of $\widehat{d}$ GPH	Mean of $\widehat{d}$ Whittle	Pr(Reject $d=0$) GPH	Pr(Reject $d=0$) Whittle	Pr(Reject $d=1$) GPH	Pr(Reject $d=1$) Whittle
0.01	0.10	0.018	0.005	0.096	0.095	0.923	0.993
	0.50	0.119	0.104	0.319	0.364	0.797	0.951
	0.80	0.458	0.410	0.834	0.872	0.569	0.764
	0.90	0.657	0.599	0.930	0.948	0.479	0.655
	0.99	0.807	0.761	0.970	0.980	0.401	0.560
0.03	0.10	0.032	0.019	0.117	0.122	0.924	0.993
	0.50	0.194	0.181	0.525	0.626	0.796	0.947
	0.80	0.539	0.498	0.957	0.981	0.553	0.718
	0.90	0.770	0.720	0.990	0.996	0.454	0.599
	0.99	0.934	0.909	0.999	1.000	0.447	0.622
0.10	0.10	0.031	0.019	0.116	0.120	0.929	0.994
	0.50	0.216	0.212	0.598	0.717	0.822	0.956
	0.80	0.539	0.532	0.989	0.998	0.501	0.649
	0.90	0.765	0.741	1.000	1.000	0.298	0.405
	0.99	0.980	0.970	1.000	1.000	0.206	0.281

Table 3: The table records estimates and tests on the long memory d for $y_t = \beta y^e_{t+1} + x_t$, under CGLS learning with gain parameter $\bar{g}$. The data is generated as $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, $T = 1000$ and the number of Monte Carlo replications is 10000. GPH is the Geweke & Porter-Hudak (1983) estimator and Whittle is the Robinson (1995) maximum local Whittle likelihood estimator. Pr (Reject $d = 0$) and Pr (Reject $d = 1$) are the empirical rejection 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.

recorded in Tables 2 and 3.

5 Application to Present Value Models

We now consider the implications of learning in present value models of stock prices and exchange rates. Specifically, we focus on the Campbell and Shiller (1987) model for stock prices, and the models of Engel and West (2005) for exchange rates. Under rational expectations, both models exhibit features that appear counterfactual and have led to the famous empirical puzzles of excess return predictability and the forward premium anomaly. Some explanations for these puzzles that have been proposed in the literature rely on the presence of long memory that is attributed to persistent shocks and is therefore of exogenous origin, see Baillie and Bollerslev (2000) and Maynard and Phillips (2001). Here, we examine whether learning can account for the persistence observed in the data even when the exogenous shocks have short memory.

There are some related papers that report results complementary to ours. Benhabib and Dave (2013) studied models for asset prices and show that some forms of learning may generate a power law for the distribution of the log dividend-price ratio. Branch and Evans (2010), and Chakraborty and Evans (2008) studied the potential of adaptive learning to explain the empirical puzzles. The former focus on explaining regime-switching in returns and their volatility, rather than low frequency properties of the dividend-price ratio, and the latter assume that fundamentals are strongly persistent.

5.1 Stock prices

Let P_t, D_t and r_t denote the price, dividend and excess return, respectively, of an index of stocks. Under the rational expectations asset pricing model of Campbell and Shiller (1988), the log dividend-price ratio is given by

$$\log \frac{D_t}{P_t} = c + E_t \sum_{j=0}^{\infty} \beta^j \left(\Delta \log D_{t+j+1} - r_{t+j+1} \right), \tag{14}$$

where c, β are log-linearization parameters, see also Campbell, Lo and McKinlay (1996, chapter 7). Equation (14) obtains as the bubble-free solution of the following first-order difference

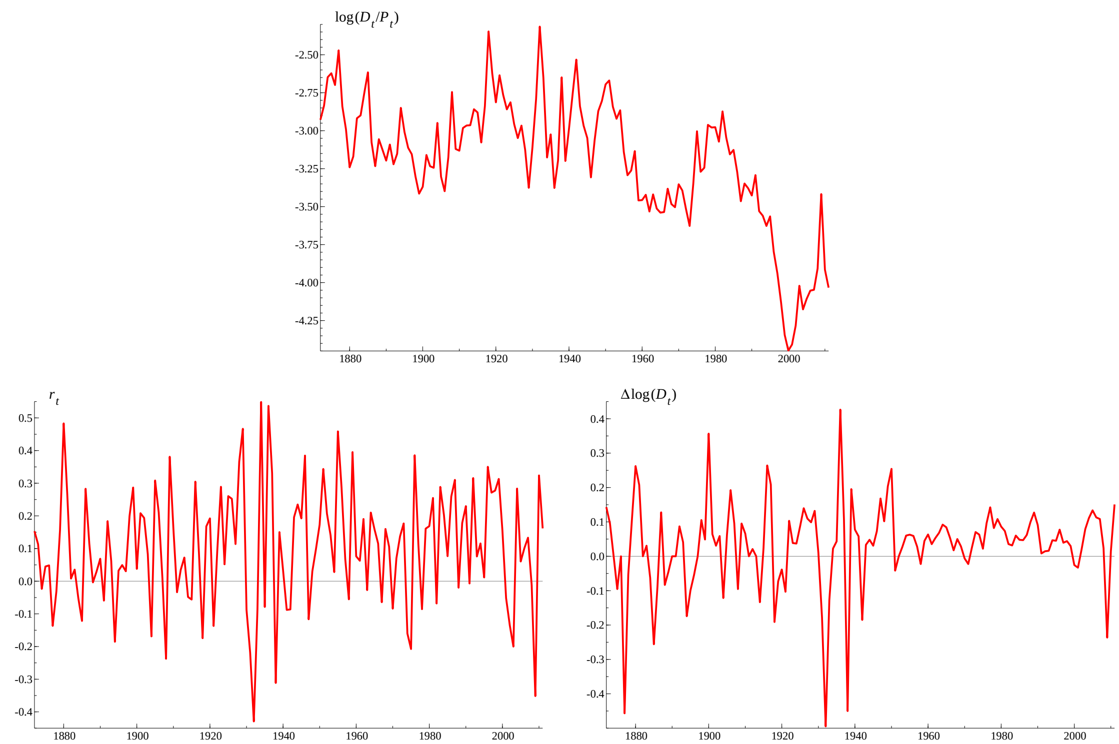

Figure 2: Log dividend-price ratio, returns and dividend growth for S&P annual index data.

equation

$$\log \frac{D_t}{P_t} = (1-\beta)\, c + \beta E_t \left(\log \frac{D_{t+1}}{P_{t+1}} \right) + E_t \left(\Delta \log D_{t+1} - r_{t+1} \right). \tag{15}$$

The above equation can be written in the form (1) with $y_t = \log \frac{D_t}{P_t}$ and $x_t = (1-\beta)\, c + E_t \left(\Delta \log D_{t+1} - r_{t+1} \right)$. We have data on y_t, but we do not observe the driving process x_t, because it depends on *expected* returns and dividend growth which are unobserved. Proposition 1 shows that if x_t exhibits short memory, then y_t should also exhibit short memory.

Figure 2 plots measures of $\log \left(D_t/P_t \right)$, r_t and $\Delta \log D_t$ using annual data on the Standard and Poor's (S&P) stock index over the period 1871-2011 available from Robert Shiller's website.[11] An apparently puzzling feature of the data is that the log dividend-price ratio exhibits very strong persistence, while dividend growth and excess returns show hardly any signs of persistence. This is demonstrated using two of the most recent estimators of the degree of memory which are both efficient and consistent under weak assumptions (Shimotsu

[11] http://www.econ.yale.edu/~shiller/data/chapt26.xls

Panel A: Stock prices and dividends

Estimator	$\log(D_t/P_t)$	r	$\Delta \log(D_t)$
2ELW	0.85	0.13	0.11
FELW	0.79	0.13	0.05
s.e.	0.15	0.15	0.15

Panel B: Forward premia

Estimator	Canada	France	Germany	Italy	Japan	UK
$\hat{d}_{2ELW}$	0.52	0.43	0.80	0.75	0.63	0.65
$\hat{d}_{FELW}$	0.50	0.50	0.80	0.68	0.63	0.50
s.e.	0.14	0.14	0.14	0.15	0.15	0.14
Sample size	151	151	151	138	137	151

Table 4: Estimates of the degree of long memory. 2ELW is the Two-Step Exact Whittle Likelihood Estimator of Shimotsu and Phillips (2005) and Shimotsu (2010), FELW is the Nonstationary-Extended local Whittle estimator of Abadir et al. (2007). Standard errors are the same for both estimators. Panel A corresponds to annual S&P data since 1871. Panel B corresponds to quarterly Eurodollar interest differentials for each of the indicated currencies from the mid-1970s.

and Phillips, 2005, Shimotsu 2010, and Abadir, Distaso and Giraitis, 2007), as reported in Panel A of Table 4. Both estimators show that y_t exhibits long memory with memory parameter 0.79 and 0.85, and significantly different from zero, while $\Delta \log D_t$ and r_t exhibit short memory.

We cannot use these empirical findings to infer that the low frequency variation in the data is inconsistent with the canonical asset pricing model for stocks under rational expectations. Specifically, an extension of an argument in Campbell, Lo and McKinlay (1996, sec. 7.1.4) can be used to show that *realized* returns and dividend growth can *appear* to exhibit short memory even though *expected* returns and/or dividend growth may have a degree of long memory that is sufficient to explain the persistence in the log dividend-price ratio. Thus, the canonical asset pricing model (14) is consistent with the observed long memory in the dividend-price ratio under rational expectations if the forcing variable x_t exhibits strong

	$\log(D_t/P_t)$	Canada	France	Germany	Italy	Japan	UK
2ELW	0.23	0.11	0.04	0.20	0.15	0.12	0.08
FELW	0.24	0.12	0.04	0.21	0.15	0.12	0.08

Table 5: The table reports the minimum value of the gain parameter such that a t-test of $\mathsf{H}_0 : d = 0$ versus $\mathsf{H}_1 : d > 0$ is not rejected for $x_t = y_t - \beta y^e_{t+1}$ at a 5% asymptotic nominal level of significance. For details of estimators and data, see Table 4.

persistence but not if x_t is a short memory process that satisfies Assumption B.

We now turn to the question of whether it is possible to explain the observed low frequency variation in $\log(D_t/P_t)$ endogenously using learning, that is, when the exogenous process x_t exhibits short memory. In our empirical analysis, we calibrate β to 0.96, based on Campbell, Lo and McKinlay (1996, chapter 7, p. 261). For any given learning algorithm, characterized by some parameter ϑ, say, we compute the expectation under learning, denoted $y^e_{t+1}(\vartheta)$, and $x_t(\vartheta) = y_t - \beta y^e_{t+1}(\vartheta)$. We then test the null hypothesis that the memory parameter, d, of $x_t(\vartheta)$ is zero against a one-sided alternative that it is positive. We use one-sided t-tests based on the Shimotsu and Phillips (2005) and Abadir *et al.* (2007) estimators, as in Table 4. If there is a value of ϑ for which the test does not reject 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 the two classes of learning algorithms studied earlier: CGLS, with $\vartheta = \bar{g} \in (0,1)$; and DGLS, with $\vartheta = \theta \in [1,5]$. Theorem 3 implies that, when β is close to one, the memory of y_t is increasing in $\bar{g}$, so we report the minimum value of $\bar{g}$ for which the null hypothesis is not rejected, *i.e.*, the minimum value of $\bar{g}$ that is consistent with the memory of y_t under CGLS learning when x_t has short memory. The results for $\log(D_t/P_t)$ are given in the first column of Table 5. Both tests yield similar values of $\bar{g} = 0.23$ and 0.24.[12] Next, we turn to DGLS algorithms covered in Theorem 2. We find that there is no value of θ for which the null hypothesis is accepted, so we conclude that DGLS learning dynamics, under the PLM considered, do not match the low frequency variation in the data.

[12] Benhabib and Dave (2013) report estimates of the gain parameter of that order of magnitude. They identify the gain though the implied tail distribution of y_t.

5.2 Exchange rates

The forward premium anomaly constitutes another puzzling empirical feature that is related to present value models and has been explained via long memory, see Maynard and Phillips (2001). The puzzle originates from the Uncovered Interest Parity (UIP) equation:

$$E_t \left[s_{t+1} - s_t\right] = f_t - s_t = i_t - i_t^* \tag{16}$$

where s_t is the log spot exchange rate, f_t is the log one-period forward rate, and i_t, i_t^* are the one-period log returns on domestic and foreign risk-free bonds and the second equality follows from the covered interest parity. The UIP under the efficient markets hypothesis has been tested since Fama (1984) as the null $\mathsf{H}_0 : (c, \gamma) = (0, 1)$ in the regression

$$\Delta s_t = c + \gamma \left(f_{t-1} - s_{t-1}\right) + \epsilon_t. \tag{17}$$

The anomaly lies in the rejection of H_0 with an estimate $\widehat{\gamma} << 1$, often negative.

Baillie and Bollerslev (2000) and Maynard and Phillips (2001) suggest econometric explanations of this puzzle that rely on strong persistence of the forward premium. Baillie and Bollerslev (2000) provide "evidence that this so-called anomaly may be viewed mainly as a statistical phenomenon that occurs because of the very persistent autocorrelation in the forward premium." Their explanation is based on persistent volatility. Maynard and Phillips (2001) show that if the forward premium $i_t - i_t^*$ is fractionally integrated and Δs_t is a short memory process that satisfies our Assumption B, then OLS estimates of γ in (17) converge to zero and have considerable probability of being negative in finite samples. They provide evidence of long memory in forward-premia for several countries relative to the US dollar. We look at the data on three-month Eurodollar interest differentials for six countries, Canada, France, Germany, Italy, Japan and the UK, over the period ranging from the mid-1970s to 2012 (starting points vary by country). The data set is the one used by Engel and West (2005), updated from Thomson Datastream.[13] Figure 3 plots the time series, and Panel B of Table 4 provides estimates of their memory parameters. We see that all series exhibit strong persistence with estimates of d greater than 0.4, corroborating the results in Maynard and Phillips (2001).

[13]Available from http://www.ssc.wisc.edu/~cengel/Data/Fundamentals/data.htm and Datastream under mnemonics S20520, S20544, S20544, S98803, S20963, S20508 and for the US: S20514.

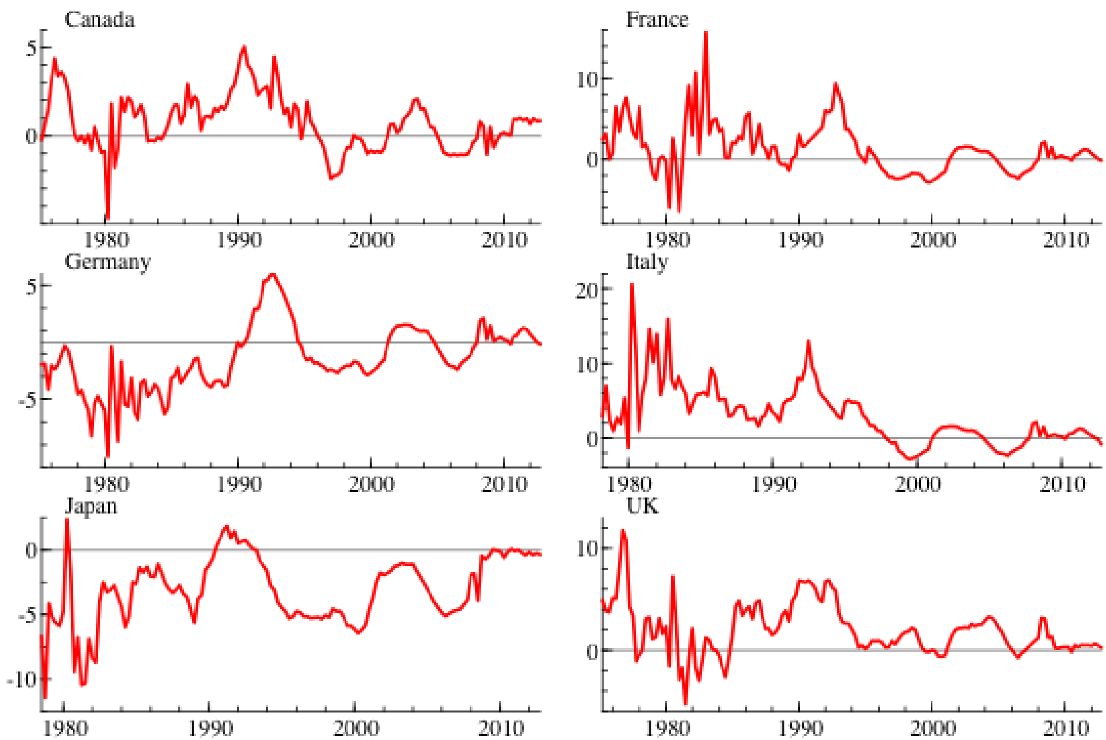

Figure 3: Forward premia with respect to the US dollar for six countries.

A possible explanation for the strong persistence in the forward premium is the presence of an exogenous time-varying risk premium, see Engel (1996). Under this explanation, the UIP equation becomes

$$E_t\left[s_{t+1} - s_t\right] = i_t - i_t^* + \rho_t, \tag{18}$$

where ρ_t is an unobserved process that represents a time-varying risk premium. In order to match the long memory of the forward premia under rational expectations, the exogenous risk premium ρ_t must exhibit long memory, too, since Δs_t appears close to *i.i.d.*, see Engel and West (2005).

We investigate whether learning dynamics can generate enough persistence to match the low frequency variation in the forward premia, without assuming that it arises exogenously through the risk premium. We consider the two exchange rate models studied in Engel and West (2005), a money-income model with an exogenous real exchange rate, and a Taylor rule model where the foreign country has an explicit exchange rate target. We show that each of these models implies a forward-looking equation for the forward premium $y_t = i_t - i_t^*$ of the form (1), with a different driving process x_t, and a different interpretation of the coefficient β for each model (derivations are given in Section I of the Appendix). Specifically, letting z_t denote a vector of 'fundamentals' that includes money, income, price and inflation differentials, the real exchange rate, and a nominal exchange rate target, it can be shown that y_t follows (1) with $x_t = (1-\beta)\left(b' E_t \Delta z_{t+1} - \rho_t\right)$, where b is a vector of coefficients that depends on the model. In the money income model, β is a function of the interest semi-elasticity of money demand, while in the Taylor rule model, β is inversely related to the degree of the intervention of foreign monetary authorities to target the exchange rate. Using past empirical studies, Engel and West (2005) calibrate β within the range $0.97 - 0.98$ for the money income model and $0.975 - 0.988$ for the Taylor rule model. For the empirical analysis here we choose the value $\beta = 0.98$, which covers both models.

We perform the same analysis as in the previous subsection, to identify any learning algorithms that can explain the persistence in y_t when x_t is short memory. The results are entirely analogous to the case of the dividend-price ratio. Specifically, we find no DGLS learning algorithm that can explain the long memory in the forward premia when the fundamentals have short memory, but we do find CGLS learning algorithms that can. The minimum gain parameters needed for each country are reported in columns 2-7 of Table 5. The smallest gain parameter corresponds to France (0.04), and the largest to Germany (0.21). These gains are somewhat higher than the values typically used in the applied learning literature, see, *e.g.*, Chakraborty and Evans (2008) for this application. All in all, our conclusions are analogous to the case of the dividend-price ratio.

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 learning can generate strong persistence. The degree of persistence induced by learning depends negatively on the weight agents place on past observations when they update their beliefs, and positively on the magnitude of the feedback from expectations to the endogenous variable. In the special case of the prototypical long-window learning algorithm known as recursive least squares, long memory arises when the coefficient on expectations is greater than a half. In algorithms with shorter window, long memory provides an approximation to the low-frequency variation of the endogenous variable. Importantly, long memory arises endogenously here, due to the self-referrential nature of the model, 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. Finally, this property of learning can be used to shed light on some well-known empirical puzzles in present value models.

Appendix

A WLS interpretation of DGLS

The model of Malmendier and Nagel (2013) assumes a non-increasing gain algorithm with gain sequence $g_t = \min(1, \theta/t)$. So, denoting by $\lceil\theta\rceil$ the *ceiling* of θ (i.e.,the smallest integer as least as large as θ),

$$y^e_{t+1} = a_t = y_{\lceil\theta\rceil} \prod_{i=0}^{t-\lceil\theta\rceil-1} (1-g_{t-i}) + \sum_{j=0}^{t-\lceil\theta\rceil-1} \left[g_{t-j} \prod_{i=0}^{j-1} (1-g_{t-i}) \right] y_{t-j}$$

$$\kappa_{t,j} = \frac{w_{t,j}}{\sum_j w_{t,j}} = \begin{cases} \frac{\theta}{t-j} \prod_{i=t-j+1}^{t} \frac{i-\theta}{i}, & \text{if } j < t - \lceil\theta\rceil; \\ 0, & j \geq t - \lceil\theta\rceil. \end{cases}.$$

Since $q(q+1)\ldots(q+n) = \frac{\Gamma(q+n+1)}{\Gamma(q)}$ if q is not a negative integer, we write

$$\kappa_{t,j} = \begin{cases} \theta \frac{\Gamma(t+1-\theta)}{\Gamma(t-j+1-\theta)} \frac{\Gamma(t-j)}{\Gamma(t+1)}, & \text{if } j < t - \lceil\theta\rceil; \\ 0, & \text{if } j \geq t - \lceil\theta\rceil. \end{cases}$$

Hence,

$$\begin{aligned} \sum_{j=1}^{t} j\kappa_{t,j} &= \theta \sum_{j=1}^{t-\lceil\theta\rceil-1} j \frac{\Gamma(t+1-\theta)}{\Gamma(t-j+1-\theta)} \frac{\Gamma(t-j)}{\Gamma(t+1)} \\ &= \theta \frac{\Gamma(t+1-\theta)}{\Gamma(t+1)} \sum_{j=\lceil\theta\rceil+1}^{t-1} (t-j) \frac{\Gamma(j)}{\Gamma(j+1-\theta)}. \end{aligned}$$

Using Stirling's fomula that for j large (see Baillie, 1996, p. 20):

$$\frac{\Gamma(j+a)}{\Gamma(j+b)} \sim j^{a-b},$$

it follows that

$$\begin{aligned} \sum_{j=1}^{t} j\kappa_{t,j} &\sim \theta t^{-\theta} \sum_{j=\lceil\theta\rceil+1}^{t-1} (t-j) j^{\theta-1} \\ &\sim \theta t^{1-\theta} \sum_{j=\lceil\theta\rceil+1}^{t-1} j^{\theta-1} - \theta t^{-\theta} \sum_{j=\lceil\theta\rceil+1}^{t-1} j^{\theta}. \end{aligned}$$

Now using $\Gamma(x+1) = x\Gamma(x)$,

$$\sum_{j=1}^{t} j\kappa_{t,j} \sim t^{1-\theta} \left[t^{\theta} - \lceil\theta\rceil^{\theta} \right] - \frac{\theta}{1+\theta} t^{-\theta} \left[t^{1+\theta} - \lceil\theta\rceil^{1+\theta} \right] \sim \frac{t}{1+\theta}.$$

So, the length of the learning window δ_κ in Definition LW is unity.

Now, $\kappa_{t,j} = \frac{w_{t,j}}{\sum_i w_{t,i}}$ and

$$\sum_{j=0}^{t} \kappa_{t,j} = \frac{\theta}{t} + \theta \frac{\Gamma(t+1-\theta)}{\Gamma(t+1)} \sum_{j=1}^{t-\lceil\theta\rceil-1} \frac{\Gamma(t-j)}{\Gamma(t-j+1-\theta)}$$
$$\sim \theta t^{-\theta} \sum_{j=\lceil\theta\rceil+1}^{t-1} j^{\theta-1} = t^{-\theta}\left(t^{\theta} - \lceil\theta\rceil^{\theta}\right)$$
$$\to 1.$$

Note that $\kappa_{t,j} = \theta \frac{\Gamma(t+1-\theta)}{\Gamma(t+1)} \frac{\Gamma(t-j)}{\Gamma(t-j+1-\theta)} \sim \theta t^{-\theta} (t-j)^{\theta-1}$ for t and $t-j$ large, with $j < t - \lceil\theta\rceil$, in which case the least-squares weights satisfy:

$$w_{t,j} = \frac{\kappa_{t,j}}{\sum_{i=0}^{t} \kappa_{t,i}} \sim \frac{\theta}{t}\left(\frac{t-j}{t}\right)^{\theta-1}.$$

B Proof of Proposition 1

We look for a solution $y_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}$ that satisfies $y_t = \beta E_t y_{t+1} + x_t$ with $\beta \leq 1$. This implies

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

Identifying the coefficients, it follows that $\psi_j - \beta\psi_{j+1} = \vartheta_j$ for all $j \geq 0$, so

$$\psi_j = \vartheta_j + \beta\psi_{j+1} = \sum_{k=j}^{\infty} \beta^{k-j} \vartheta_k.$$

Hence as $j \to \infty$, $\psi_j \to 0$ and the rate of decay of the (ψ_j) coefficients will be slowest when $\beta = 1$. When $\beta < 1$,

$$\psi_j = O\left(\vartheta_j \sum_{k=0}^{\infty} \beta^k\right) = O(\vartheta_j),$$

so $\left|\sum_{j=0}^{\infty} \psi_j\right| < \infty$ if $\left|\sum_{j=0}^{\infty} \vartheta_j\right| < \infty$. We then use Theorem 3.11 of Phillips and Solo (1992) who show that y_t then satisfies a Central Limit Theorem.

If $\beta = 1$, then $\psi_j = \sum_{k=j}^{\infty} \vartheta_k$ so

$$\sum_{j=0}^{\infty} \psi_j = \sum_{j=0}^{\infty} (j+1) \vartheta_j$$

and the result follows from the assumption $\sum_{j=0}^{\infty} j |\vartheta_j| < \infty$.

C Derivation of expression (10)

Substituting $y_{t+1}^{e}=\frac{1}{t}\sum_{s=1}^{t}y_{s}$ into (1) yields

$$y_{t}=\beta\frac{1}{t}\sum_{s=1}^{t}y_{s}+x_{t}=\frac{t}{t-\beta}x_{t}+\frac{\beta}{t-\beta}\sum_{s=1}^{t-1}\prod_{k=s}^{t-1}\frac{k}{k-\beta}x_{s}.$$

Hence,

$$\frac{\partial y_{t+j}}{\partial x_{t}}=\frac{\beta}{t+j-\beta}\prod_{k=t}^{t+j-1}\frac{k}{k-\beta}=\beta\frac{\Gamma\left(t+j\right)\Gamma\left(t-\beta\right)}{\Gamma\left(t\right)\Gamma\left(t+j+1-\beta\right)}$$

so using Stirling's formula, as $(t,j/t)\rightarrow(\infty,\infty)$,

$$\begin{aligned}\frac{\partial y_{t+j}}{\partial x_{t}}&\sim\beta t^{-\beta}\left(t+j\right)^{-(1-\beta)}=\beta t^{-1}\left(1+j/t\right)^{-(1-\beta)}\\&\sim\beta t^{-\beta}j^{-(1-\beta)}.\end{aligned}$$

D Proof of Theorem 2

In the proof of the theorem, we make use of the following lemma that derives the rate of decay of the autocovariance of an Assumption B process x_t, and is an extension of a result mentioned in Hosking (1996) whose proof is in Hosking (1994, p. 5). Hosking's result is for $a\in(1/2,1)$ but we show that the result holds also for $a>1$, a not an integer.

Lemma 5 *If $x_{t}=\sum_{j=0}^{\infty}\vartheta_{j}\epsilon_{t-j}$, where $\vartheta_{j}\sim\delta j^{-a}$, $\delta>0$, $a>1$, $a\notin\mathbb{N}$, and ϵ_t is white noise with finite variance σ_{ϵ}^{2} then $\gamma_{x}\left(j\right)\sim c_{x}j^{1-2a}$, where $c_x>0$ is a constant.*

Proof. Assume that a is not an integer. For $k\geq0$, let

$$\Psi_{k}=\delta\frac{\Gamma\left(k+1-a\right)}{\Gamma\left(k+1\right)}$$

which is defined by continuity when $k+1-a<0$ and $|k+1-a|\notin\mathbb{N}$. Stirling's formula implies that as $k\rightarrow\infty$, $\Psi_{k}\sim\delta k^{-a}$. Then (i) there exists $C>0$ such that $|\vartheta_{k}|\leq C\Psi_{k}$ for all $k\geq0$; also (ii) for all $\epsilon>0$, there exists K such that for $k\geq K$, $\vartheta_{k}\in\left(\left(1-\epsilon\right)\Psi_{k},\left(1+\epsilon\right)\Psi_{k}\right)$.

Then,

$$\sum_{k=0}^{\infty}\Psi_{k}\Psi_{k+j}=\delta^{2}\sum_{j=0}^{\infty}\frac{\Gamma\left(k+1-a\right)}{\Gamma\left(k+1\right)}\frac{\Gamma\left(k+j+1-a\right)}{\Gamma\left(k+j+1\right)}$$

which can be expressed in terms of the hypergeometric function $F(1-a, j+1-a; j+1; 1)$, which is defined for these parameter values, since $j+1 > (1-a) + (j+1-a)$. Using Gauss's theorem for the value at $F(\cdot,\cdot;\cdot;1)$ expressed in terms of Gamma functions, we obtain as $j \to \infty$,

$$\sum_{k=0}^{\infty} \Psi_k \Psi_{k+j} = \delta^2 \frac{\Gamma(1-a)\Gamma(2a-1)\Gamma(j+1-a)}{\Gamma(a)\Gamma(j+a)}$$
$$\sim \delta^2 \frac{\Gamma(1-a)\Gamma(2a-1)}{\Gamma(a)} j^{1-2a} = \mathcal{O}\left(j^{1-2a}\right).$$

where the second line follows using Stirling's formula. Now $\gamma_x(j) = \sigma_\epsilon^2 \sum_{k=0}^{\infty} \vartheta_k \vartheta_{k+j}$ and

$$\left| j^{2a-1} \left(\gamma_x(j) - \sigma_\epsilon^2 \sum_{k=0}^{\infty} \Psi_k \Psi_{k+j} \right) \right|$$
$$\leq \sigma_\epsilon^2 j^{2a-1} \left(\sum_{k=0}^{K-1} |\vartheta_k \vartheta_{k+j}| + \sum_{k=0}^{K-1} \Psi_k \Psi_{k+j} + \sum_{k=K}^{\infty} |\vartheta_k \vartheta_{k+j} - \Psi_k \Psi_{k+j}| \right)$$
$$\leq \sigma_\epsilon^2 j^{2a-1} K \left(C^2+1\right) \Psi_0 \Psi_j + \sigma_\epsilon^2 j^{2a-1} \left(2\epsilon + \epsilon^2\right) \sum_{k=K}^{\infty} \Psi_k \Psi_{k+j}$$

Hence, since $\sigma_\epsilon^2 j^{2a-1} K \left(C^2+1\right) \Psi_0 \Psi_j = O\left(j^{a-1}\right)$, there exist $M > 0$ such that as $j \to \infty$

$$\left| j^{2a-1} \left(\gamma_x(j) - \sigma_\epsilon^2 \sum_{k=0}^{\infty} \Psi_k \Psi_{k+j} \right) \right| \leq M\epsilon$$

so

$$\gamma_x(j) \sim \delta^2 \frac{\Gamma(1-a)\Gamma(2a-1)}{\Gamma(a)} j^{1-2a}.$$

■

Proof of the theorem 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 (6a), $a_t = \frac{1-g_t}{1-\beta g_t} a_{t-1} + \frac{g_t}{1-\beta g_t} x_t$ or

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

with $\prod_{j=t+1}^{t} \left(1 - \frac{(1-\beta) g_j}{1-\beta g_j}\right) \equiv 1$. 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[14]

$$a_t^* = \prod_{j=1}^{t} (1 - (1-\beta) g_j) a_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} (1 - (1-\beta) g_j) g_i x_i. \quad (19)$$

[14] In the specific situation where $g_1 = 1$ the impact of a_0 on a_t is zero contrary to that on a_t^*. This only concerns g_1 since $g_{i+1} < g_i \leq 1$ for all $i \geq 1$; it does not affect the magnitude of $\mathsf{Var}[S_T]$ as we show later.

Hence, we can infer the order of magnitude of $\mathsf{Var}\,(S_T)$ from that of $\mathsf{Var}\,(S_T^*)$, where $S_T^* = \sum_{t=1}^{T} (\beta a_t^* + x_t)$. Using (19), S_T^* can be written as

$$S_T^* = \beta h_{T+1} a_0 + \sum_{t=1}^{T} \phi_{T,t} x_t,$$

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

For clarity, we first consider the case when x_t is serially uncorrelated, and treat the general case at the end. The variance of S_T^* is given by

$$\mathsf{Var}\,[S_T^*] = \beta^2 h_{T+1}^2 \mathsf{Var}\,(a_0) + \sigma_x^2 \sum_{t=1}^{T} \phi_{T,t}^2, \tag{20}$$

where $\sigma_x^2 = \mathsf{Var}\,[x_t]$. We study each of the two terms on the right hand side of the above expression.

The asymptotic rates of h_t and k_t depend on the value of $(1-\beta)\,\theta$. Since $g_i \sim \theta i^{-1}$, $g_i^2 = o(g_i)$. We first assume $(1-\beta)\,\theta \neq 0$. Then for i large enough so $(1-\beta)\, g_i < 1$, $\log(1 - (1-\beta) g_i) = -(1-\beta) g_i + o(g_i)$ and $\log k_t = -(1-\beta)\,\theta \log t + o(\log t)$. Thus, $g_t/k_t \sim \theta t^{-1}/t^{-(1-\beta)\theta} = \theta t^{(1-\beta)\theta - 1}$. Turning to $h_t = \sum_{i=1}^{t-1} k_i$,

$$h_t \sim \begin{cases} t^{1-(1-\beta)\theta} / \left[1 - (1-\beta)\,\theta\right], & \text{if } (1-\beta)\,\theta < 1; \\ \log t, & \text{if } (1-\beta)\,\theta = 1; \\ \zeta\left((1-\beta)\,\theta\right), & \text{if } (1-\beta)\,\theta > 1, \end{cases} \tag{21}$$

where $\zeta(u)$ is Riemann's zeta function evaluated at $u > 1$ (the case $\beta = 0$ is included for completeness, since it plays no role in the asymptotic rates of $\mathsf{Var}\,(S_T^*)$). It follows that as $t \to \infty$, for $t \le T$,

$$\phi_{T,t} \sim \begin{cases} 1 + \frac{\beta\theta}{1-(1-\beta)\theta} \left(\left(\frac{T}{t}\right)^{1-(1-\beta)\theta} - 1 \right), & \text{if } (1-\beta)\,\theta < 1; \\ 1 + \beta\theta \log \frac{T}{t}, & \text{if } (1-\beta)\,\theta = 1; \\ 1 + \beta\theta t^{(1-\beta)\theta - 1} \left(\sum_{i=t}^{T+1} i^{-(1-\beta)\theta} \right), & \text{if } (1-\beta)\,\theta > 1, \end{cases} \tag{22}$$

Consider first $(1-\beta)\,\theta < 1$ so

$$\phi_{T,t}^2 \sim \left[\frac{\beta\theta}{1 - (1-\beta)\,\theta} \right]^2 \left(\frac{T}{t} \right)^{2[1-(1-\beta)\theta]}$$

The second term in variance of S_T^*, see eq. (20), is:

$$\sum_{t=1}^{T}\phi_{T,t}^2 \sim \begin{cases} \left[\frac{\beta\theta}{1-(1-\beta)\theta}\right]^2 \zeta\left(2\left[1-(1-\beta)\theta\right]\right) T^{2[1-(1-\beta)\theta]}, & \text{if } (1-\beta)\theta < \frac{1}{2}; \\ 4(1-\theta)^2 T\log T, & \text{if } (1-\beta)\theta = \frac{1}{2}; \\ \left[\frac{1-\theta}{1-(1-\beta)\theta}\right]^2 \frac{1}{1-2[1-(1-\beta)\theta]} T, & \text{if } \frac{1}{2} < (1-\beta)\theta < 1. \end{cases}$$

Now, if $(1-\beta)\theta = 1$, then

$$\begin{aligned}\sum_{t=1}^{T}\phi_{T,t}^2 &\sim \left[\frac{\beta\theta}{1-(1-\beta)\theta}\right]^2 \left[T\log^2 T - 2\log T\left(T\log T - T\right) + T\left(\log^2 T - 2\log T + 2\right)\right] \\ &= 2\left[\frac{\beta\theta}{1-(1-\beta)\theta}\right]^2 T\end{aligned}$$

Finally, if $(1-\beta)\theta > 1$, then $\phi_{T,t} \sim 1 + \beta\theta t^{(1-\beta)\theta-1}\left(\sum_{i=t}^{T+1} i^{-(1-\beta)\theta}\right)$, where

$$1 \leq 1 + \beta\theta t^{(1-\beta)\theta-1}\left(\sum_{i=t}^{T+1} i^{-(1-\beta)\theta}\right) \leq \frac{1-\theta}{1-(1-\beta)\theta} + \frac{\beta\theta}{1-(1-\beta)\theta}\left(\frac{T+1}{t}\right)^{1-(1-\beta)\theta}.$$

Hence, since

$$\sum_{t=1}^{T}\left[\frac{1-\theta}{1-(1-\beta)\theta}\left(\frac{T+1}{t}\right)^{1-(1-\beta)\theta}\right]^2 = \mathcal{O}(T) \tag{23}$$

it follows that $\sum_{t=1}^{T}\phi_{T,t}^2 = \mathcal{O}(T)$. Summarizing,

$$\sum_{t=1}^{T}\phi_{T,t}^2 = \begin{cases} \mathcal{O}\left(T^{2[1-(1-\beta)\theta]}\right), & \text{if } (1-\beta)\theta < \frac{1}{2}; \\ \mathcal{O}(T\log T), & \text{if } (1-\beta)\theta = \frac{1}{2}; \\ \mathcal{O}(T), & \text{if } (1-\beta)\theta > \frac{1}{2}. \end{cases} \tag{24}$$

Next, we examine the order of magnitude of the first term in eq. (20), $h_{T+1}^2 \mathsf{Var}(a_0)$. First, note that:

$$h_{T+1}^2 = \begin{cases} \mathcal{O}\left(T^{2[1-(1-\beta)\theta]}\right), & \text{if } (1-\beta)\theta < 1; \\ \mathcal{O}\left(\log^2 T\right), & \text{if } (1-\beta)\theta = 1; \\ \mathcal{O}(1), & \text{if } (1-\beta)\theta > 1. \end{cases}$$

Combining with the assumption that $a_0 = O_p(1)$, the contribution of $h_{T+1}^2\mathsf{Var}(a_0)$ to $T^{-1}\mathsf{Var}[S_T^*]$ is asymptotically negligible when $(1-\beta)\theta \geq 1/2$. When $(1-\beta)\theta < 1/2$, $T^{-1}h_{T+1}^2\mathsf{Var}(a_0) = \mathcal{O}\left(T^{1-2(1-\beta)\theta}\right)$. The result of the theorem then follows from the rates in (24).

When $(1-\beta)\,\theta = 0$, which only arises if $\beta = 1$, then θ is irrelevant for the magnitude so we set it to unity. To ensure a proper definition of the learning algorithm we let $y_2^e = y_1 - x_t$, i.e., $\kappa_{1,0} = 1$ and $\varphi_1 = -x_1$ in 7, so for $t \geq 2$,

$$a_t = y_1 + \sum_{i=2}^{t} \frac{x_i}{i-1}$$

and

$$S_T = y_1 T + \frac{T}{T-1} x_T + T \sum_{t=2}^{T-1} \frac{x_t}{t-1}. \tag{25}$$

Hence expression (22) extends to the case where $\beta = 1$, and as $T \to \infty$

$$\mathsf{sd}\left(T^{-1/2} S_T\right) = \mathcal{O}\left(T^{1/2}\right).$$

Now, we turn to the general case where x_t is not serially uncorrelated, and denote by $\gamma_x(\cdot)$ its autocovariance function. Then $\mathsf{Var}\,(S_T^*)$ contains the following term, in addition to the two terms in eq. (20):

$$2 \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} \gamma_x(i). \tag{26}$$

First, we use Lemma 5 to characterize the rate of decay of $\gamma_x(j)$. Assumption B imposes that $\sum_{j=0}^{\infty} j\,|\vartheta_j| < \infty$, so there exists $a > 2$, $a \notin \mathbb{N}$, such that $|\vartheta_j| = O\left(j^{-a}\right)$ Hence, $\gamma_x(j) = O\left(j^{-\varpi}\right)$ for $\varpi = 2a - 1 > 3$, and there exist $c_x > 0$ such that

$$|\gamma_x(j)| \leq c_x j^{-\varpi}.$$

Now, consider

$$\left| \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} \gamma_x(i) \right| \leq c_x \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} i^{-\varpi}.$$

It suffices to establish that $\sum_{i=1}^{T-t} \phi_{T,t+i} i^{-\varpi} = O\left(\phi_{T,t}\right)$. Observe that

$$\sum_{i=1}^{T-t} \phi_{T,t+i} i^{-\varpi} \sim \begin{cases} \sum_{i=1}^{T-t} i^{-\varpi} + \frac{\beta\theta}{1-(1-\beta)\theta} \sum_{i=1}^{T-t} \left(\left(\frac{T}{t+i}\right)^{1-(1-\beta)\theta} - 1 \right) i^{-\varpi}, & \text{if } (1-\beta)\,\theta < 1; \\ \sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=1}^{T-t} i^{-\varpi} \log \frac{T}{t+i}, & \text{if } (1-\beta)\,\theta = 1; \\ \sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=1}^{T-t} (t+i)^{(1-\beta)\theta - 1} \left(\sum_{j=t+i}^{T+1} j^{-(1-\beta)\theta} \right) i^{-\varpi}, & \text{if } (1-\beta)\,\theta > 1, \end{cases}$$

Consider first the case $(1-\beta)\,\theta > 1$. Then,

$$\sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=1}^{T-t} (t+i)^{(1-\beta)\theta-1} \left(\sum_{j=t+i}^{T+1} j^{-(1-\beta)\theta} \right) i^{-\varpi} \tag{27}$$
$$\in \left[\sum_{i=1}^{T-t} i^{-\varpi}, \sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=1}^{T-t} (t+i)^{(1-\beta)\theta-1} \left(\sum_{j=t+i}^{T+1} j^{-(1-\beta)\theta} \right) i^{-\varpi} \right]$$

with $\sum_{j=t+i}^{T+1} j^{-(1-\beta)\theta} \le \int_{t+i-1}^{T+1} u^{-(1-\beta)\theta} du = ((1-\beta)\,\theta-1)^{-1} \left[(t+i-1)^{1-(1-\beta)\theta} - (T+1)^{1-(1-\beta)\theta} \right]$. So expression (27) is bounded below by $\frac{1-(T-t+1)^{1-\varpi}}{\varpi-1}$ and above by

$$\frac{1-(T-t+1)^{1-\varpi}}{\varpi-1} + \frac{\beta\theta}{(1-\beta)\,\theta-1} \sum_{i=1}^{T-t} (t+i)^{(1-\beta)\theta-1} \left((t+i-1)^{1-(1-\beta)\theta} - (T+1)^{1-(1-\beta)\theta} \right) i^{-\varpi}$$
$$\le \frac{1-(T-t+1)^{1-\varpi}}{\varpi-1} + \frac{\beta\theta}{(1-\beta)\,\theta-1} \left(\sum_{i=1}^{T-t} i^{-\varpi} - (T+1)^{1-(1-\beta)\theta} \sum_{i=1}^{T-t} (t+i)^{(1-\beta)\theta-1}\, i^{-\varpi} \right)$$
$$\le \frac{1-(T-t+1)^{1-\varpi}}{\varpi-1} + \frac{\beta\theta}{(1-\beta)\,\theta-1} \left[\frac{1}{\varpi-1} \right]$$
$$- \left(\frac{(T-t+1)^{1-\varpi}}{\varpi-1} + \frac{(T+1)^{1-(1-\beta)\theta}\,(T-t)^{(1-\beta)\theta-\varpi}}{(1-\beta)\,\theta-\varpi} - \frac{(T+1)^{1-(1-\beta)\theta}}{(1-\beta)\,\theta-\varpi} \right)$$

Now, since $\varpi > 1$ and $(1-\beta)\,\theta > 1$,

$$\frac{(T-t+1)^{1-\varpi}}{\varpi-1} = O(1), \quad \frac{(T+1)^{1-(1-\beta)\theta}}{(1-\beta)\,\theta-\varpi} = O(1),$$

and also

$$\left| \frac{(T+1)^{1-(1-\beta)\theta}\,(T-t)^{(1-\beta)\theta-\varpi}}{(1-\beta)\,\theta-\varpi} \right| \le \frac{(T-t)^{1-\varpi}}{|(1-\beta)\,\theta-\varpi|} = O(1).$$

Hence,

$$\sum_{i=1}^{T-t} \phi_{T,t+i} i^{-\varpi} = \mathcal{O}(1). \tag{28}$$

If $(1-\beta)\,\theta = 1$, then

$$\sum_{i=1}^{T-t} \phi_{T,t+i} i^{-\varpi} \sim \sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=1}^{T-t} i^{-\varpi} \log \frac{T}{t+i}$$
$$= \sum_{i=1}^{T-t} i^{-\varpi} + \beta\theta \sum_{i=0}^{T-(t+1)} (T-i)^{-\varpi} \log \frac{T}{T+t-i},$$

where, as T become large,

$$(T-i)^{-\varpi}\log\frac{T}{T+t-i}\sim-(T-i)^{-\varpi}\,\frac{t-i}{T}.$$

So,

$$\begin{aligned}
-\sum_{i=0}^{T-(t+1)}(T-i)^{-\varpi}\,\frac{t-i}{T} &= \sum_{i=t+1}^{T} i^{-\varpi}\frac{T-t-i}{T}\\
&\leq \frac{T-t}{t}\left|\frac{T^{1-\varpi}-t^{1-\varpi}}{1-\varpi}\right|+T^{-1}\left|\frac{T^{2-\varpi}-t^{2-\varpi}}{2-\varpi}\right|\\
&= O\left(T^{2-\varpi}\right)=O\left(1\right),
\end{aligned}$$

since $\varpi>2$. Thus, (28) holds.

When $(1-\beta)\,\theta<1$ (including the case $\beta=1$), it suffices to show that

$$\sum_{i=1}^{T-t}\phi_{T,t+i}\gamma_x\left(i\right)=O\left(\left(\frac{T}{t}\right)^{1-(1-\beta)\theta}\right),$$

i.e.,

$$\sum_{i=1}^{T-t}\phi_{T,t+i}i^{-\varpi}=O\left(t^{(1-\beta)\theta-1}\right). \tag{29}$$

Substituting for $\phi_{T,t+i}$ from (22), and ignoring the constants, we have

$$\begin{aligned}
&\sum_{i=1}^{T-t}\left(\left(\frac{T}{t+i}\right)^{1-(1-\beta)\theta}-1\right)i^{-\varpi}\\
&=T^{1-(1-\beta)\theta}\sum_{i=1}^{T-t}(t+i)^{(1-\beta)\theta-1}\,i^{-\varpi}-\sum_{i=1}^{T-t}i^{-\varpi}.
\end{aligned}$$

The last term is bounded since $\varpi>3$, so we focus on the first term. Using the hypergeometric function F, we have

$$\begin{aligned}
&\sum_{i=1}^{T-t}(t+i)^{(1-\beta)\theta-1}\,i^{-\varpi}\leq\int_1^{T-t+1}(t+i)^{(1-\beta)\theta-1}\,i^{-\varpi}di\\
&=\frac{t^{(1-\beta)\theta-1}}{1-\varpi}\left[u^{1-\varpi}F\left(1-\varpi,1-\left(1-\beta\right)\theta;2-\varpi;-u/t\right)\right]_{u=1}^{u=T-t+1}\\
&\sim\frac{t^{(1-\beta)\theta-1}}{\varpi-1}-\frac{t^{(1-\beta)\theta-1}}{\varpi-1}\left(T-t+1\right)^{1-\varpi}F\left(1-\varpi,1-\left(1-\beta\right)\theta;2-\varpi;-\frac{T-t+1}{t}\right)\\
&\leq\frac{t^{(1-\beta)\theta-1}}{\varpi-1},
\end{aligned}$$

where the last two steps follow from the fact that $t \leq T$ and $\varpi > 1$ (in fact, $\varpi > 3$ here), and $F(1-\varpi, 1-(1-\beta)\theta; 2-\varpi; z) \to 1$ as $z \to 0$ and remains bounded when $z \to -\infty$. This establishes (29), and the result in the theorem follows.

E Proof of expression (12)

Under CGLS learning the algorithm is

$$\kappa_t(L) = \bar{g} \sum_{j=0}^{t-1} (1-\bar{g})^j L^j,$$

and $\varphi_t = a_0 (1-\bar{g})^t$. Hence

$$\begin{aligned} m(\kappa_t) &= \bar{g} \sum_{j=1}^{t-1} j (1-\bar{g})^j = -\bar{g}(1-\bar{g}) \frac{\partial}{\partial \bar{g}} \sum_{j=0}^{t-1} (1-\bar{g})^j \\ &= (1-\bar{g}) \frac{1-(1-\bar{g})^{t-1}[1+(t-1)\bar{g}]}{\bar{g}}. \end{aligned}$$

Now consider $m(\kappa_T)$, and assume that $\bar{g} = c_g T^{-\lambda}$. Then $(1-\bar{g})^{T-1} = \exp\left\{(T-1)\log\left(1-c_g T^{-\lambda}\right)\right\}$ and as $T \to \infty$

$$(1-\bar{g})^{T-1} \sim \exp\left\{-c_g \frac{T-1}{T^{\lambda}}\right\} \to \begin{cases} 0, & \text{if } \lambda < 1; \\ e^{-c_g}, & \text{if } \lambda = 1. \end{cases}$$

Turning to the mean lag, for $\lambda < 1$ $(1-\bar{g})^{t-1}[1+(t-1)\bar{g}] \to 0$ so $m(\kappa_T) \sim \frac{T^{\lambda}}{c_g}$. When $\lambda = 1$, $m(\kappa_T) \sim \frac{1-e^{-c_g}[1+c_g]}{c_g} T$, which proves (12).

F Proof of Theorem 3

Under the stated assumptions, the estimator a_t is generated by

$$a_t = \frac{\bar{g}}{1-\beta\bar{g}} \sum_{i=1}^{t} \left(1 - \frac{(1-\beta)\bar{g}}{1-\beta\bar{g}}\right)^{t-i} x_i.$$

When β is local to unity and $\bar{g}$ local to zero, $1 - \frac{(1-\beta)\bar{g}}{1-\beta\bar{g}} \sim 1-(1-\beta)\bar{g}$, so we define

$$a_t^* = \bar{g} \sum_{i=1}^{t} (1-(1-\beta)\bar{g})^{t-i} x_i,$$

which is simpler to analyze using existing results.

Define $\xi_t = \bar{g}^{-1} a_t^*$ such that

$$\xi_t = \sum_{i=1}^{t} \left(1 - (1-\beta)\,\bar{g}\right)^{t-i} x_i,$$

with $(\beta, \bar{g}) = \left(1 - c_\beta T^{-\nu}, c_g T^{-\lambda}\right)$ for $(\nu, \lambda) \in [0,1]^2$. Several cases arise depending on the values of λ, ν. These correspond to a_t exhibiting an exact unit root for $\bar{g} = 0$ or $\beta = 1$, a near-unit root for $\lambda + \nu = 1$ (see Chan and Wei, 1987, and Phillips 1987), a moderate-unit root for $\lambda + \nu \in (0,1)$ (see Giraitis and Phillips, 2006, Phillips and Magdalinos, 2007 and Phillips, Magdalinos and Giraitis, 2010) and a very-near-unit root for $\lambda + \nu > 1$ (see Andrews and Guggenberger, 2007). Under x_t satisfying Assumption B, their results imply:

$$\xi_T = \begin{cases} O_p(1), & \lambda = \nu = 0; \\ O_p\left(T^{(\lambda+\nu)/2}\right), & \lambda + \nu \in (0,1); \\ O_p\left(T^{1/2}\right), & \lambda + \nu \geq 1. \end{cases}$$

Also $\frac{(1-\beta)\bar{g}}{1-\beta\bar{g}} = \mathcal{O}\left((1-\beta)\,\bar{g}\right)$ implies that $S_T^* = \sum_{t=1}^{T} \beta a_t^* + x_t = \mathcal{O}_p\left(\sum_{t=1}^{T} \beta a_t + x_t\right)$. To derive the magnitude of $S_T^* = \beta\bar{g} \sum_{t=1}^{T} \xi_{t-1} + \sum_{t=1}^{T} x_t$ we notice that:

$$\sum_{t=1}^{T} \xi_t = \sum_{t=1}^{T} \sum_{i=1}^{t} \left(1 - (1-\beta)\,\bar{g}\right)^{t-i} x_i = \sum_{t=1}^{T} \frac{1 - \left(1 - (1-\beta)\,\bar{g}\right)^{T-t+1}}{1 - \left(1 - (1-\beta)\,\bar{g}\right)} x_t,$$

i.e.,

$$\sum_{t=1}^{T} \xi_t = \frac{1}{(1-\beta)\,\bar{g}} \left[\sum_{t=1}^{T} x_t - \left(1 - (1-\beta)\,\bar{g}\right)\xi_T\right].$$

Hence

$$\bar{g} \sum_{t=1}^{T} \xi_t = \frac{1}{(1-\beta)} \left(\sum_{t=1}^{T} x_t - \xi_T\right) + \bar{g}\xi_T. \tag{30}$$

We start with the case $\nu + \lambda < 1$, where $\xi_T = o\left(\sum_{t=1}^{T} x_t\right)$. Expression (30) implies that $\bar{g} \sum_{t=1}^{T} \xi_t = O_p\left(T^{1/2+\nu}\right)$ and hence

$$\mathsf{sd}\left(T^{-1/2} S_T^*\right) = \mathcal{O}\left(T^{\nu}\right).$$

If $\nu + \lambda = 1$, then Phillips (1987) – see also Stock (1994, example 4, p. 2754) – shows that

$$T^{-1/2} \left(\sum_{t=1}^{T} x_t - \xi_T\right) = T^{-1/2} \sum_{i=1}^{T} \left(1 - \left(1 - (1-\beta)\,\bar{g}\right)^{T-i}\right) x_i$$
$$\Rightarrow \int_0^1 \left(1 - e^{-c_\beta c_g (1-r)}\right) dW(r) = O_p(1),$$

where $T^{-1/2}\sum_{t=1}^{\lceil rT\rceil} x_t \Rightarrow W(r)$, where $W(\cdot)$ is a Brownian motion and $\Rightarrow$ denotes weak convergence of the associated probability measure. It follows that $\sum_{t=1}^{T} x_t - \xi_T = O\left(T^{1/2}\right)$ and expression (30) implies that $\bar{g}\sum_{t=1}^{T}\xi_t = O_p\left(T^{1/2+\nu}\right)$. Hence

$$\mathsf{sd}\left(T^{-1/2}S_T^*\right) = \mathcal{O}\left(T^{\nu}\right) = \mathcal{O}\left(T^{1-\lambda}\right).$$

Now, if $\nu + \lambda > 1$,

$$\begin{aligned}\sum_{t=1}^{T} x_t - \xi_T &= \sum_{i=0}^{T-1}\left[1-(1-(1-\beta)\bar{g})^i\right]x_{T-i}\\ &= ((1-\beta)\bar{g})\sum_{i=0}^{T-1}\left[i+\mathcal{O}\left(i^2((1-\beta)\bar{g})\right)\right]x_{T-i}.\end{aligned}$$

It is well known that $\sum_{i=0}^{T-1} i x_{T-i} = O_p\left(T^{3/2}\right)$ and $\sum_{i=0}^{T-1} i^2 x_{T-i} = O_p\left(T^{5/2}\right)$ (see, *e.g.*,Hamilton 1994, chap. 17). Hence $(1-\beta)\bar{g}\sum_{i=0}^{T-1} i^2 x_{T-i} = o\left(\sum_{i=0}^{T-1} i x_{T-i}\right)$, and, in expression (30):

$$\frac{1}{(1-\beta)}\left(\sum_{t=1}^{T} x_t - \xi_T\right) + \bar{g}\xi_T = O_p\left(T^{3/2-\lambda}\right) + O_p\left(T^{1/2-\lambda}\right).$$

When $\lambda < 1$, $3/2 - \lambda > 1/2$ so $\sum_{t=1}^{T} x_t = o_p\left(\bar{g}\sum_{t=1}^{T}\xi_{t-1}\right)$, and the order of magnitude of S_T^* follows from that of $\bar{g}\sum_{t=1}^{T}\xi_{t-1}$:

$$\mathsf{sd}\left(T^{-1/2}S_T^*\right) = \mathcal{O}\left(T^{1-\lambda}\right).$$

If $\lambda = 1$, $\sum_{t=1}^{T} x_t = O_p\left(\bar{g}\sum_{t=1}^{T}\xi_{t-1}\right)$ and the previous expression also applies.

G Proof of Theorem 4

We introduce the following two lemmas which we use in the proof. These are proven in a supplementary appendix.

Lemma 6 *Let $\kappa(L) = \sum_{j=0}^{\infty}\kappa_j L^j$ with $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$ as $j \to \infty$, for $c_\kappa > 0$ and $\delta_\kappa \in (0,1)$. Assume $\kappa(1) = 1$. Then, there exist $c_\kappa^* \neq 0$ and $c_\kappa^{**} > 0$ such that*

$$\begin{aligned}&\mathrm{Re}\left(\kappa\left(e^{i\omega}\right)-1\right) \underset{\omega\to 0^+}{=} -c_\kappa^*\omega^{1-\delta_\kappa} + o\left(\omega^{1-\delta_\kappa}\right),\\ &\left|\kappa\left(e^{i\omega}\right)-1\right|^2 \underset{\omega\to 0^+}{=} c_\kappa^{**}\omega^{2(1-\delta_\kappa)} + o\left(\omega^{2(1-\delta_\kappa)}\right).\end{aligned}$$

Proof. see the supplementary appendix. ■

Lemma 7 *Consider the model* $y_t = \beta y^e_{t+1} + x_t$*, with* $y^e_{t+1} = \kappa(L) y_t$*. Suppose* x_t *satisfies Assumption B, and that the constant learning algorithm* $\kappa(\cdot)$ *satisfies Assumption A with* $\delta_\kappa \in (0,1)$*. We assume that* β *is fixed and* $\beta < \kappa(1)$*. Then the spectral density of* y_t *is finite at the origin* $f_y(0) < \infty$ *and admits an upward vertical asymptote: there exists* $c_f > 0$ *such that*

$$f_y'(0) \underset{\omega\to 0}{\sim} -c_f \omega^{-\delta_\kappa}. \tag{31}$$

Proof. See the supplementary appendix. ■

In the proof of Theorem 4, we omit for notational ease the dependence of β, the spectral densities and autocovariances on T (this is particularly important when referring to Lemma 7).

Substitute (7) into (1) to get

$$y_t = \beta \sum_{j=0}^{t-1} \kappa_j y_{t-j} + \beta \varphi_t + x_t,$$

and define $\kappa^*(L) = 1 - \kappa(L) = \sum_{j=0}^{\infty} \kappa_j^* L^j$ so

$$(1-\beta) y_t + \beta \sum_{j=0}^{t-1} \kappa_j^* y_{t-j} = x_t + \beta \varphi_t.$$

Summing yields

$$\sum_{t=1}^{T} \left((1-\beta) - \beta \sum_{j=0}^{t-1} \kappa_j^* \right) y_{T-t+1} = \sum_{t=1}^{T} (x_t + \beta \varphi_t). \tag{32}$$

The left-hand side of the previous equation shows that the magnitude of $\sum_{t=1}^{T} y_t$ depends on the limit of $(1-\beta) / \sum_{j=0}^{T-1} \kappa_j^*$. Since $\kappa^*(1) = 0$, if there exists $\lambda < 1$ such that $\kappa_j \sim c_\kappa j^{\lambda-2}$ then $\kappa_j^* \sim -c_\kappa j^{\lambda-2}$ and $\sum_{j=0}^{T-1} \kappa_j^* \sim \frac{c_\kappa}{1-\lambda} T^{\lambda-1}$. Under Assumption A, the previous expressions hold letting $\lambda = \delta_\kappa$ when $\delta_\kappa \in (0,1)$; when $\delta_\kappa = 0$, there exists $\lambda < 0$ such that $\kappa_j = O\left(j^{\lambda-2}\right)$ and $\kappa_j^* = O\left(j^{\lambda-2}\right)$ since Assumption A.3 rules out $\kappa_j \sim c_\kappa j^{-2}$.

Let $\beta = 1 - c_\beta T^{-\nu}$. Defining $y_t^- = y_t 1_{\{t \le 0\}}$, we made the following assumptions about φ_t :

$$\begin{cases} \varphi_t = \kappa(L) y_t^-, & \text{if } \delta_\kappa \in \left(\frac{1}{2}, 1\right); \\ \Delta\varphi_t = (1-L)\kappa(L) y_t^-, & \text{if } \delta_\kappa \in \left(0, \frac{1}{2}\right). \end{cases} \tag{33}$$

so $(1-\beta\kappa(L))\,y_t = x_t$ if $\delta_\kappa \in (1/2,1)$ or $(1-\beta\kappa(L))\,\Delta y_t = \Delta x_t$ if $\delta_\kappa \in (0,1/2)$. Hence $(1-\beta\kappa(1))\,E(y_t) = E(x_t)$ or $(1-\beta\kappa(1))\,E(\Delta y_t) = E(\Delta x_t)$ so the random variables y_t, x_t can be expressed in deviation from their expectations. In other words, we may assume without loss of generality and for ease of exposition that $E(x_t) = 0$ since this does not affect the variances and spectral densities.

Consider the case $\nu > 1-\delta_\kappa$ so $(1-\beta)/\sum_{j=0}^{T-1}\kappa_j^* \to 0$. This rules out $\delta_\kappa = 0$. First assume that $\delta_\kappa \in \left(\frac{1}{2},1\right)$. Define $z_t = [\kappa^*(L)]^{-1}\,x_t$ with spectral density

$$f_z(\omega) = \frac{f_x(\omega)}{|1-\kappa(e^{-i\omega})|^2}.$$

Using lemma 6, with $c_\kappa^{**} > 0$, as $\omega \to 0$

$$f_z(\omega) \sim \frac{f_x(0)}{c_\kappa^{**}}\omega^{-2(1-\delta_\kappa)}. \tag{34}$$

Beran (1994, theorem 2.2 p. 45) shows that (34) implies that

$$\mathsf{Var}\left(\sum_{t=1}^{T} z_t\right) = \mathcal{O}\left(T^{1+2(1-\delta_\kappa)}\right).$$

The proof is in the appendix of Beran (1989) and relies on showing that $f_z(\omega)$ can be written as $\left|1-e^{-i\omega}\right|^{-2(1-\delta_\kappa)} S(1/\omega)$ where S is slowly varying at infinity.

Under assumption (33), noting that $\kappa(L)\,y_t^- = (\kappa(L)-1)\,y_t^-$, expression (32) rewrites

$$\sum_{t=1}^{T}\left((1-\beta)-\beta\sum_{j=0}^{t-1}\kappa_j^*\right)y_{T-t+1} - \beta\sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t} = \sum_{t=1}^{T} x_t.$$

Since $(1-\beta) = o\left(\sum_{j=0}^{T-1}\kappa_j^*\right)$, it follows that, denoting $y_t^+ = y_t - y_t^-$,

$$\begin{aligned}
&\sum_{t=1}^{T}\left((1-\beta)-\beta\sum_{j=0}^{t-1}\kappa_j^*\right)y_{T-t+1} - \beta\sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t} \\
&= -\beta\left[\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1}\kappa_j^*\right)y_{T-t+1} + \sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t}\right] + o_p\left(\sum_{t=1}^{T}\sum_{j=0}^{t-1}\kappa_j^* y_{T-t+1}\right) \\
&= \sum_{t=1}^{T}(1-\kappa(L))\,y_t + o_p\left(\sum_{t=1}^{T}(1-\kappa(L))\,y_t^+\right).
\end{aligned}$$

Hence, using $\sum_{t=1}^{T} x_t = \sum_{t=1}^{T} (1-\kappa(L)) z_t$,

$$\begin{aligned}\sum_{t=1}^{T} (1-\kappa(L)) y_t + o_p\left(\sum_{t=1}^{T} (1-\kappa(L)) y_t^{+}\right) &= \sum_{t=1}^{T} x_t \\ \sum_{t=1}^{T} (1-\kappa(L)) (y_t - z_t) + o_p\left(\sum_{t=1}^{T} (1-\kappa(L)) y_t^{+}\right) &= 0 \\ \sum_{t=1}^{T} (y_t - z_t) + o_p\left(\sum_{t=1}^{T} y_t\right) &= 0\end{aligned}$$

i.e.

$$\sqrt{\mathsf{Var}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)} = \mathcal{O}\left(T^{1-\delta_\kappa}\right). \tag{35}$$

Now, if $\delta_\kappa \in (0, 1/2)$, defining $\Delta z_t = [\kappa^*(L)]^{-1} \Delta x_t$, and following the previous steps starting from $(1-\beta\kappa(L))\Delta y_t = \Delta x_t$ leads to

$$\sum_{t=1}^{T} \Delta(y_t - z_t) + o_p\left(\sum_{t=1}^{T} \Delta y_t\right) = 0.$$

The result by Beran (1989) regarding the magnitude of $\mathsf{Var}\left(\sum_{t=1}^{T} \Delta z_t\right)$ cannot be used here for $(1-\delta_\kappa) \in \left(\frac{1}{2}, 1\right)$. Yet, the spectral density of Δz_t satisfies

$$f_{\Delta z}(\omega) \sim \frac{f_x(0)}{c_\kappa^{**}} \omega^{2\delta_\kappa},$$

which implies (see Lieberman and Phillips, 2008) that there exists $c_\gamma \neq 0$ such that $\gamma_{\Delta z}(k) \sim c_\gamma k^{-2\delta_\kappa - 1}$. Also $f_{\Delta z}(0) = 0$ so $\gamma_{\Delta z}(0) + 2\sum_{k=1}^{\infty} \gamma_{\Delta z}(k) = 0$. The long run variance of Δz_t is hence such that

$$\begin{aligned}\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T} \Delta z_t\right) &= \gamma_{\Delta z}(0) + 2T^{-1}\sum_{k=1}^{T-1} (T-k)\gamma_{\Delta z}(k) \\ &= \left(\gamma_{\Delta z}(0) + 2\sum_{k=1}^{T-1} \gamma_{\Delta z}(k)\right) - 2T^{-1}\sum_{k=1}^{T-1} k\gamma_{\Delta z}(k) \\ &= -\sum_{k=T}^{\infty} \gamma_{\Delta z}(k) - 2T^{-1}\sum_{k=1}^{T-1} k\gamma_{\Delta z}(k) \\ &= \mathcal{O}\left(T^{-2\delta_\kappa}\right). \end{aligned} \tag{36}$$

We now consider the case $\nu \leq 1-\delta_\kappa$, starting with assuming $\delta_\kappa \neq 0$ so $\nu < 1$. Brillinger (1975, theorem 5.2.1) shows that if the covariances of y_t are summable,

$$\frac{\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T} y_t\right)}{f_y(0)} = (2\pi T)^{-1}\int_{-\pi}^{\pi}\frac{\sin^2(T\omega/2)}{\sin^2(\omega/2)}\frac{f_y(\omega)}{f_y(0)}d\omega, \tag{37}$$

where $f_y(\omega)$ is the spectral density of y_t. The function $\left[\frac{\sin(T\omega/2)}{\sin(\omega/2)}\right]^2$ achieves its maximum over $[-\pi,\pi]$ at zero where its value is T^2. As $T\to\infty$ it remains bounded for all $\omega\neq 0$. It is therefore decreasing in ω in a neighborhood of 0^+. For any given T, Lemma 7 – for β fixed – shows that $f_y(\omega)$ is also decreasing in such a neighborhood and $\frac{f_y(\omega)}{f_y(0)}$ is bounded. Both functions in the integrand of (37) being positive, their product is also decreasing in ω in a neighborhood of 0^+; it is in addition continuous, even and differentiable at all $\omega\neq 0$. As $T\to\infty$, the integrand of (37) presents a pole at the origin and its behavior in the neighborhood of zero governs the magnitude of the integral. Since the integrand achieves its local maximum at zero, we can restrict our analysis to a neighborhood thereof, $[0,\theta_T]$ with $\theta_T = o\left(T^{-1}\right)$ since $\frac{\sin^2(T\theta_T/2)}{\sin^2(\theta_T/2)}\frac{f_y(\omega)}{f_y(0)}$ remains bounded as $T\to\infty$ for any sequence θ_T such that $T\theta_T \nrightarrow 0$.

Let $\varepsilon > 0$ and $\beta = 1 - c_\beta T^{-\nu}$, we develop the integrand of (37) about the origin, provided $T^\nu\theta_T^{1-\delta_\kappa} = \left(T^{\nu/(1-\delta_\kappa)}\theta_T\right)^{1-\delta_\kappa} = o(1)$, i.e., if $\nu \leq 1-\delta_\kappa$. This yields for the integral over $[0,\theta_T]$:

$$\begin{aligned}
&(2\pi T)^{-1}\int_0^{\theta_T}\left(T^2\left(1-\frac{1}{3}\left(T^2-1\right)\omega^2+o\left(T^2\omega^2\right)\right)\right)\left(1-c_V T^\nu\omega^{1-\delta_\kappa}+o\left(T^\nu\omega^{1-\delta_\kappa}\right)\right)d\omega\\
&=\frac{T}{2\pi}\left[\theta_T-\frac{1}{9}\left(T^2-1\right)\theta_T^3-\frac{c}{2-\delta_\kappa}T^\nu\theta_T^{2-\delta_\kappa}+\frac{c_V}{3(4-\delta_\kappa)}\left(T^2-1\right)T^\nu\theta_T^{4-\delta_\kappa}\right]\\
&=\frac{T}{2\pi}\left[T^{-(1+\varepsilon)}-\frac{T^2-1}{9}T^{-3(1+\varepsilon)}-\frac{c_V}{2-\delta_\kappa}T^{\nu-(2-\delta_\kappa)(1+\varepsilon)}+\frac{c_V\left(T^2-1\right)}{3(4-\delta_\kappa)}T^\nu T^{-(4-\delta_\kappa)(1+\varepsilon)}\right]\\
&\sim\frac{1}{2\pi}\left[T^{-\varepsilon}-\frac{1}{9}T^{-3\varepsilon}-\frac{c_V}{2-\delta_\kappa}T^{\nu-(1-\delta_\kappa)-(2-\delta_\kappa)\varepsilon}+\frac{c_V}{3(4-\delta_\kappa)}T^{\nu-(1-\delta_\kappa)-(4-\delta_\kappa)\varepsilon}\right],
\end{aligned}\tag{38}$$

where c_V is implicitly defined from Lemma 7. Expression (38) shows that if $\nu\leq 1-\delta_\kappa$ the integral over $[0,\theta_T]$ – and hence that over $[-\pi,\pi]$ – remains bounded in the neighborhood of the origin and hence $\frac{\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T}y_t\right)}{f_y(0)} = O(1)$, with $f_y(0) = (1-\beta)^{-2}f_x(0) = \mathcal{O}\left(T^{2\nu}\right)$. Hence $\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T}y_t\right) = O\left(T^{2\nu}\right)$ and

$$\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T}y_t\right) = \mathcal{O}\left(T^{2\nu}\right). \tag{39}$$

Finally, when $(\delta_\kappa, \nu) = (0, 1)$, Assumption A.3 implies that $0 < \kappa'(1) = \sum_{j=1}^{\infty} j\kappa_j < \infty$. By Lemma 2.1 of Phillips and Solo (1992), there exists a polynomial $\widetilde{\kappa}$ such that

$$\kappa(L) = 1 - (1 - L)\,\widetilde{\kappa}(L),$$

with $\widetilde{\kappa}(1) < \infty$. $\widetilde{\kappa}(L) = (1 - L)^{-1}(1 - \kappa(L))$ so the roots of $\widetilde{\kappa}$ coincide with the values z such that $\kappa(z) = 1$, except at $z = 1$ for which $\widetilde{\kappa}(1) = \kappa'(1) > 0$ (by L'Hospital's rule and assumption A.3). $\kappa(z) = 1$ and $c_\kappa > 0$ together imply that the roots of $\widetilde{\kappa}(L)$ lie outside the unit circle ($\kappa(z) < \kappa(1) = 1$ for $|z| \leq 1$, $z \neq 1$) and the process $\widetilde{x}_t$ defined by $\widetilde{\kappa}(L)\,\widetilde{x}_t = x_t$ is $\mathsf{I}(0)$ with differentiable spectral density at the origin by Assumption B (Stock, 1994, p. 2746). Hence y_t satisfies the near-unit root definition of Phillips (1987):

$$(1 - \beta L)\, y_t = \widetilde{x}_t,$$

and the result follows from Stock (1994, example 4 p. 2754) since $\widetilde{x}_t$ satisfies his conditions (2.1)-(2.3).

H Alternative definitions of memory parameter for the algorithm with hyperbolic weights

The following definitions of the memory parameter d, are equivalent to (9) for covariance stationary processes, see Beran (1994) or Baillie (1996):

$$\begin{aligned} \rho_z(k) &\sim c_\rho k^{2d-1}, && \text{as } k \to \infty \\ f_z(\omega) &\sim c_f\,|\omega|^{-2d}, && \text{as } \omega \to 0, \end{aligned} \tag{40}$$

for some positive constants c_ρ, c_f, where $\rho_z(k) = \mathsf{Corr}[z_t, z_{t+k}]$ is the autocorrelation function (ACF) of a covariance stationary stochastic process z_t and $f_z(\omega)$ is its spectral density. For $d > 0$, the autocorrelation function at long lags and the spectrum at low frequencies have the familiar hyperbolic shape that has traditionally been used to define long memory.

Fractional integration, denoted I(d), is a well-known example of a class of processes that exhibit long memory. When $d < 1$, the process is mean reverting (in the sense of Campbell and Mankiw, 1987, that the impulse response function to fundamental innovations converges to zero, see Cheung and Lai, 1993). Moreover, I(d) processes admit a covariance stationary representation when $d \in (-1/2, 1/2)$, and are non-stationary if $d \geq 1/2$. Long memory arises when the degree of fractional integration is positive, $d > 0$. In the case of nonstationary processes, the ACF definition of d in (40) does not apply,[15] so we use the ACF/spectrum of Δz, as in Heyde and Yang (1997):

$$\begin{aligned} \rho_{\Delta z}(k) &\sim c_\rho k^{2(d-1)-1}, \quad 1/2 < d < 1 \qquad \text{as } k \to \infty; \\ f_{\Delta z}(\omega) &\sim c_f |\omega|^{-2(d-1)}, \quad 1/2 < d < 1 \qquad \text{as } \omega \to 0. \end{aligned} \tag{41}$$

We prove the following theorem in the supplementary appendix.

Theorem 8 *Under the assumptions of Theorem 4 where the spectral density of x_t has bounded second order derivative, if $\nu > 1 - \delta_\kappa$, then:*
1. the spectral density f_y of y_t evaluated at Fourier frequencies $\omega_j = 2\pi j/T$ with $j = 1, ...n$, and $n = o(T)$, satisfies as $T \to \infty$,

$$f_y(\omega_j) \sim f_x(0)\, \omega_j^{-2(1-\delta_\kappa)}$$

2. the autocorrelation functions ρ_y of y_t, or $\rho_{\Delta y}$ of Δy_t, evaluated at $k = o(T)$, satisfy as $T, k \to \infty$,

$$\begin{aligned} \rho_y(k) &= \mathcal{O}\left(k^{1-2\delta_\kappa}\right) \quad \text{if } \tfrac{1}{2} < \delta_\kappa < 1 \\ \rho_{\Delta y}(k) &= \mathcal{O}\left(k^{-2\delta_\kappa - 1}\right) \quad \text{if } 0 < \delta_\kappa < \tfrac{1}{2}. \end{aligned}$$

The theorem shows that the degree of memory measured in Theorem 4 through Definition LM coincides with common alternative definitions.

I Derivation of models for the forward premium

We derive expression (1) for $y_t = i_t - i_t^*$ from the money-income and Taylor rule models of Engel and West (2005). We show below that both of these models imply a relationship between the log spot exchange rate s_t and y_t of the form

$$s_t = \alpha y_t + b' z_t, \tag{42}$$

[15]The property $f_z(\omega) \sim c_f |\omega|^{-2d}$ can be applied also to nonstationary cases with $1/2 < d < 1$ if $f_z(\omega)$ is defined in the sense of Solo (1992) as the limit of the expectation of the sample periodogram.

where z_t consists of price, money, income, inflation, output gap money demand shock and policy shock differentials, and the real exchange rate, and b is a vector of parameters that is derived below for each model. Substituting in the UIP equation (18) and re-arranging yields

$$\begin{aligned} s_t + y_t &= E_t s_{t+1} - \rho_t \\ (1+\alpha)\, y_t + b' z_t &= \alpha E_t y_{t+1} + b' E_t z_{t+1} - \rho_t \\ y_t &= \frac{\alpha}{1+\alpha} E_t y_{t+1} + \frac{1}{1+\alpha} \left[b' E_t \Delta z_{t+1} - \rho_t \right]. \end{aligned}$$

This is in the form (1) with $\beta = \frac{\alpha}{1+\alpha}$ and $x_t = (1-\beta)\left[b' E_t \Delta z_{t+1} - \rho_t\right]$.

Now, we derive (42) for each of the two models in Engel and West (2005).

Money-income model The money market relationship for the home country (Engel and West, 2005, Equation (4) on p. 492) is given by

$$m_t = p_t + \gamma y_t - \alpha i_t + v_{mt}, \tag{43}$$

where m_t is the log of the home money supply, p_t is the log of the home price level, i_t is the level of the home interest rate, y_t is the log of output, and v_{mt} is a shock to money demand. A similar relationship holds for the foreign country with variables $m_t^*, p_t^*, y_t^*, i_t^*$ and v_{mt}^*, and identical coefficients α and γ. The nominal exchange rate is given by

$$s_t = p_t - p_t^* + q_t \tag{44}$$

where q_t is the (exogenous) real exchange rate (Engel and West, 2005, Equation (5) on p. 493). Subtracting the foreign from the home money market relationship yields

$$p_t - p_t^* = m_t - m_t^* + \gamma\left(y_t^* - y_t\right) + v_{mt}^* - v_{mt} + \alpha\left(i_t - i_t^*\right).$$

Substituting this into (44) yields (42) with $y_t = i_t - i_t^*$ and

$$b' z_t = m_t - m_t^* + \gamma\left(y_t^* - y_t\right) + v_{mt}^* - v_{mt} + q_t.$$

Taylor rule model Suppose the home country follows the Taylor rule (Engel and West, 2005, Equation (9) on p. 494)

$$i_t = \beta_1 y_t^g + \beta_2 \pi_t + v_t, \tag{45}$$

where $\pi_t = p_t - p_{t-1}$ and y_t^g is the "output gap". The foreign country follows the Taylor rule (Engel and West, 2005, Equation (10) on p. 494)

$$i_t^* = -\beta_0 \left(s_t - \bar{s}_t^*\right) + \beta_1 y_t^{*g} + \beta_2 \pi_t^* + v_t^*, \tag{46}$$

where $\beta_0 \in (0, 1)$ and $\bar{s}_t^*$ is the target for the exchange rate. Assume further that $\bar{s}_t^* = p_t - p_t^*$ (the Purchasing Power Parity level of the exchange rate), see Engel and West (2005, Equation (11) on p. 495). Subtracting (46) from (45) yields

$$i_t - i_t^* = \beta_0 s_t - \beta_0 \left(p_t - p_t^*\right) + \beta_1 \left(y_t^g - y_t^{*g}\right) + \beta_2 \left(\pi_t - \pi_t^*\right) + \left(v_t - v_t^*\right).$$

Re-arranging the above equation yields (42) with $y_t = i_t - i_t^*$, $\alpha = 1/\beta_0$, and

$$b' z_t = \left(p_t - p_t^*\right) - \frac{\beta_1}{\beta_0}\left(y_t^g - y_t^{*g}\right) - \frac{\beta_2}{\beta_0}\left(\pi_t - \pi_t^*\right) - \frac{1}{\beta_0}\left(v_t - v_t^*\right).$$

References

Abadir, K. M., W. Distaso, and L. Giraitis (2007). Nonstationarity-extended local Whittle estimation. *Journal of Econometrics 141*, 1353–1384.

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

Andrews, D. W. K. and P. Guggenberger (2007). Asymptotics for stationary very nearly unit root processes. *Journal of Time Series Analysis 29*(1), 203–212.

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.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Baillie, R. T. and T. Bollerslev (2000). The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance 19*, 471488.

Benhabib, J. and C. Dave (2013). Learning, large deviations and rare events. *Review of Economic Dynamics* (forthcoming).

Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika 76*(2), pp. 261–269.

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

Berenguer-Rico, V. and J. Gonzalo (2013). Summability of stochastic processes (a generalization of integration and co-integration valid for non-linear processes). *Journal of Econometrics* (forthcoming).

Bobkoski, M. (1983). Hypothesis testing in nonstationary time series. Unpublished PhD thesis, Dept. of Statistics, University of Wisconsin, Madison.

Branch, W. and G. W. Evans (2010). Asset return dynamics and learning. *Review of Financial Studies 23*, 1651–80.

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory*. New York: Holt, Rinehart and Winston. Reprinted in 2001 as a SIAM Classic in Applied Mathematics.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1996). *The Econometrics of Financial Markets*. London: Princeton University Press.

Campbell, J. Y. and N. G. Mankiw (1987). Are output fluctuations transitory? *Quarterly Journal of Economics 102*(4), 857–880.

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

Campbell, J. Y. and R. J. Shiller (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies 1*(3), 195–228.

Chakraborty, A. and G. W. Evans (2008). Can perpetual learning explain the forward premium puzzle? *Journal of Monetary Economics 55*, 477–90.

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

Chan, N. H. and C. Z. Wei (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics 15*(3), 1050–1063.

Cheung, Y.-W. and K. S. Lai (1993). A fractional cointegration analysis of purchasing power parity. *Journal of Business and Economic Statistics 11*(1), 103–112.

Chevillon, G., M. Massmann, and S. Mavroeidis (2010). Inference in models with adaptive learning. *Journal of Monetary Economics 57*(3), 341–51.

Clarida, R., J. Galí, and M. Gertler (1999). The Science of Monetary Policy: A New Keynesian Perspective. *Journal of Economic Literature 37*(4), 1661–1707.

Davidson, J. and N. Hashimzade (2008). Alternative frequency and time domain versions of fractional Brownian motion. *Econometric Theory 24*(1), 256–293.

Davidson, J. and N. Hashimzade (2009). Type I and type II fractional Brownian motions: a reconsideration. *Computational Statistics and Data Analysis 53*(6), 2089–2106.

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.

Diebold, F. X. and G. D. Rudebusch (1991). On the power of dickey-fuller tests against fractional alternatives. *Economics Letters 35*(2), 155 – 160.

Durbin, J. and S. J. Koopman (2008). *Time Series Analysis by State Space Methods*. Oxford University Press. 2nd ed.

Engel, C. and K. D. West (2005). Exchange rates and fundamentals. *Journal of Political Economy 113*(3), 485–517.

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.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics 14*(3), 319–338.

Giraitis, L. and P. C. B. Phillips (2006). Uniform limit theory for stationary autoregression. *Journal of Time Series Analysis 27*, 51–60.

Gonzalo, J. and J.-Y. Pitarakis (2006). Threshold effects in cointegrating relationships. *Oxford Bulletin of Economics and Statistics 68*, 813–833.

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. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics 48*(3), 213–228.

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

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

Heyde, C. C. and Y. Yang (1997). On defining long range dependence. *Journal of Applied Probability 34*, 939–944.

Johansen, S. (2008). Fractional autoregressive processes. *Econometric Theory 24*, 651–676.

Lieberman, O. and P. C. B. Phillips (2008). A complete asymptotic series for the autocovariance function of a long memory process. *Journal of Econometrics 147*(1), 99 – 103.

Magdalinos, T. and P. C. B. Phillips (2009). Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors. *Econometric Theory 25*, 482–526.

Malmendier, U. and S. Nagel (2013). Learning from inflation experiences. Working paper, UC Berkeley.

Marcet, A. and T. J. Sargent (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory 48*, 337368.

Marinucci, D. and P. M. Robinson (1999). Alternative forms of fractional brownian motion. *Journal of Statistical Planning and Inference 80*(1), 111–122.

Maynard, A. and P. C. B. Phillips (2001). Rethinking an old empirical puzzle: Econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics 16*(6), 671–708.

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.

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.

Perron, P. and Z. Qu (2010). Long-memory and level shifts in the volatility of stock market return indices. *Journal of Business and Economic Statistics 28*, 275–290.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika 74*(3), 535–547.

Phillips, P. C. B. (2007). Regression with slowly varying regressors and nonlinear trends. *Econometric Theory 23*, 557–614.

Phillips, P. C. B. and T. Magdalinos (2007). Limit theory for moderate deviations from a unit root. *Journal of Econometrics 136*, 115–130.

Phillips, P. C. B., T. Magdalinos, and L. Giraitis (2010). Smoothing local-to-moderate unit root theory. *Journal of Econometrics 158*(2), 274–79.

Phillips, P. C. B. and V. Solo (1992). Asymptotics for linear processes. *Annals of Statistics 20*(2), 971–1001.

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.

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.

Stock, J. H. (1994). Unit roots, structural breaks and trends. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 46, pp. 2739–2841. Elsevier.

Tanaka, K. (1999). The nonstationary fractional unit root. *Econometric Theory 15*, 549–82.

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.