Learning, Large Deviations and Rare Events*

Jess Benhabib† NYU

Chetan Dave‡ NYU

Revised: September 23, 2012

Abstract

We examine the role of generalized constant gain stochastic gradient (SGCG) learning in generating large deviations of an endogenous variable from its rational expectations value. We show analytically that these large deviations can occur with a frequency associated with a fat tailed distribution even though the model is driven by thin tailed exogenous stochastic processes. We characterize these large deviations that are driven by sequences of consistently low or consistently high shocks. We then apply our model to the canonical asset-pricing model. We demonstrate that the tails of the stationary distribution of the price-dividend ratio will follow a power law.

Keywords: Adaptive learning, large deviations, fat tails, asset prices

JEL Codes: D80, D83, D84

*We thank Chryssi Giannitsarou, In-Koo Cho, John Duffy, George Evans, Boyan Jovanovic, Tomasz Sadzik, Benoite de Saporta, Tom Sargent and two anonymous referees for helpful comments and suggestions. The usual disclaimer applies.

†Address: New York University, Department of Economics, 19 W. 4th Street, 6FL, New York, NY, 10012, USA. E-mail: jess.benhabib@nyu.edu. Telephone: 1-646-567-7316, Fax: 1-212-995-4186

‡Address: New York University, Department of Economics, 19 W. 4th Street, 6FL, New York, NY, 10012, USA. E-mail: cdave@nyu.edu.

1. Introduction

Dynamic stochastic models have at times difficulty matching some features of macroeconomic data.[1] One route to reconcile differences between data and theory has been to replace the assumption of rational expectations with that of adaptive learning, where agents are assumed to estimate the underlying parameters of a model via recursive least squares. For example, if the monetary authority adaptively learns the underlying Phillips curve via decreasing gain least squares regressions, then the Nash outcome is the one that is selected (see Evans and Honkapohja (2001)). Still, the U.S. economy escaped the high inflation of the 1970's predicted by the standard model with decreasing gains. To provide an explanation Sargent (1999) and Cho et al. (2002) assumed instead that a monetary authority estimates a misspecified Phillips curve using constant gain algorithms that place more weight on recent observations. This assumption allowed the possibility of escape from a Nash outcome to a low inflation (Ramsey) outcome. In particular, within the context of their endogenous tracking model, a sequence of otherwise rare shocks can cause frequent large deviations from a high inflation self-confirming equilibrium. Indeed Sargent et al. (2006) take these endogenous tracking models to the data and account for the behavior of inflation in the U.S.

Our analysis also focuses on the role of large deviations theory and its interplay with constant gain learning dynamics. Specifically, working within the adaptive learning tradition set out by Sargent and Williams (2005), Evans et al. (2010) and others, we examine the role of generalized constant gain stochastic gradient (SGCG) learning algorithms in generating

[1] For example, empirical evaluations of consumption based asset pricing models lead to numerous asset pricing puzzles, and evaluations of real business cycle models cannot typically account for the pattern of hours worked without appealing to labor supply elasticities that are often at odds with microeconometric evidence.

large deviations of an endogenous variable from its rational expectations value. We show analytically that these large deviations can occur with a frequency associated with a fat tailed distribution even though the model is driven by thin tailed exogenous stochastic processes. Using some new techniques in the analysis of stochastic processes and linear recursions with multiplicative noise[2], we characterize these large deviations occurring under adaptive learning that are driven by sequences of consistently low or consistently high shocks. Such sequences are rare in that the average of realizations in the sequences can significantly diverge from the population mean of the shocks. We then apply our model to the single asset version of the canonical model of Lucas (1978) that has been studied extensively by Carceles-Poveda and Giannitsarou (2007, 2008) who look at the ability of learning models to approximate the behavior of aggregate stock market data.

A particular issue in the modification of standard rational expectations models to better account for features of the data by introducing adaptive learning is the choice of the learning algorithm itself. Typically, in replacing the rational expectations assumption with that of adaptive learning, agents are assumed to estimate parameters of processes to be forecasted using recursive (adaptive) methods.[3] A particular strain of this literature demonstrates the consistency of this approach with Bayes' Law. In a stationary model with optimal learning, estimated parameters ultimately converge to their rational expectations equilibrium. In recent work however, Sargent and Williams (2005) introduce a model where agents expect a random walk drift in estimated parameters. They then show that the SGCG algorithm,

[2]See Kesten (1973), Saporta (2005) and Roiterstein (2007).

[3]In asset pricing contexts, see for example: Adam et al. (2008), Adam and Marcet (2011), Branch and Evans (2010), Brennan and Xia (2001), Bullard and Duffy (2001), Carceles-Poveda and Giannitsarou (2008), Cogley and Sargent (2008), and Timmermann (1993, 1996).

that assigns more weight to recent observations on account of the underlying drift in the estimated parameters, is asymptotically the optimal Bayesian estimator. Evans et al. (2010) follow Sargent and Williams (2005) and show how a SGCG learning algorithm approximates an optimal (in a Bayesian sense) Kalman filter. Under such adaptive SGCG learning, uncertainty about estimated parameters persists over time and can fuel escape dynamics in which a sequence of consistently high or consistently low shocks propel agents away from the REE of a model.[4] In an asset-pricing context Weitzman (2007) also shows that if recent observations are given more weight under Bayesian learning of the variance of the consumption growth rate, agents will forecast returns and asset prices using thick-tailed distributions for consumption growth.[5] It is for this reason that we focus on an asset pricing context to analytically demonstrate how SGCG learning, consistent with optimal Bayesian learning, can account for the data features and fat tailed distributions of the price-dividend ratio.

Theoretically, we demonstrate that under adaptive learning of the asset prices, the tails of the stationary distribution of the price-dividend ratio will follow a power law, even though the dividend process has thin tails and is specified as a stationary $AR(1)$ process. The tail index or power-law coefficient of the price-dividend ratio can be expressed as a function of model parameters, and in particular of the optimal gain parameter that assigns decaying weights to older observations. In fact, as demonstrated by Sargent and Williams (2005)

[4]See also Holmstrom (1999) for an application to managerial incentives of learning with an underlying drift in parameters.

[5]See also Koulovatianos and Wieland (2011). They adopt the notion of rare disasters studied by Barro (2009) in a Bayesian learning environment. They find that volatility issues are well addressed. Similarly Chevillon and Mavroeidis (2011) find that giving more weight to recent observations under learning can generate low frequency variability observed in the data. See also Gabaix (2009) who provides an excellent summary of instances in which economic data follow power laws and suggests a number of causes of such laws for financial returns. In particular, Gabaix et al. (2006) suggest that large trades in illiquid asset markets on the part of institutional investors could generate extreme behavior in trading volumes (usually predicted to be zero in Lucas-type environments) and returns.

and more recently by Evans et al. (2010), the optimal gain depends on the variance of the underlying drift in the estimated parameters: the higher the variance of the drift parameter, the higher the gain, and the thicker the tail of the distribution of the price-dividend ratio. We characterize how the power law tail index of the of the long-run stationary distribution of the price-dividend ratio varies as a function of the gain parameter and of the other deep parameters of the model. Under our adaptive learning scheme that approximates optimal Bayesian learning, stationary dividend processes generate distributions for the price-dividend ratio that are not Normal. Thus, large deviations of the price-dividends ratio from the rational expectations equilibrium are possible with a frequency higher than that associated with a Normal distribution even though the dividend process is thin-tailed.

Our analysis and simulations indicate that under standard parameter calibrations, to match either the empirical tail index or the variance of the annual "fat-tailed" price dividend ratio, we require a gain parameter around 0.4-0.55, significantly higher than what is typically used in the adaptive learning literature (0.01-0.04). Carceles-Poveda and Giannitsarou (2008) also employ large parameter values for the gain in asset pricing contexts, as do Branch and Evans (2010). The latter implicitly assumes slowly decaying weights on past observations, and therefore very little underlying drift in the parameters estimated by agents. In order to get an empirical handle on the gain parameter we estimate the parameters of our model, including the gain parameter, by two separate methods. The first is a structural minimum distance estimation method for the tail index. This method puts higher weight on the empirically observed tail of the price-dividend ratio, and produces a gain estimate in the range of 0.35-0.53. The second method computes the gain as Bayesian agents expecting drifting parameters would, using a Kalman filter on the data. This yields a gain parameter

in the range of 0.49-0.55, assigning decaying weights on past observations that take the parameter drift into account. Therefore agents who use this gain parameter would indeed have their expectations confirmed by the data.

The paper is structured as follows. We first describe the general dynamic stochastic equation under learning, and also briefly illustrate its application to the single asset pricing version of Lucas (1978). Then in Section 3 we prove that our learning model, written as a random linear recursion with multiplicative noise, predicts that the tails of the stationary distribution of the endogenous variable of interest, in our application the price-dividends ratio, will follow a power law with coefficient κ that is a function of model parameters. In Section 4 we use simulations to study how κ varies with the deep parameters. In Section 5 we provide estimates of the deep parameters of the model for our asset pricing application, and of the gain parameter in particular, that are consistent with the κ estimated directly from the price-dividends ratio. Section 6 concludes.

2. Model Environment

We focus on models of the type

$$p_t = \beta E_t(p_{t+1}) + \theta d_t \tag{1}$$

in which the exogenous driving process d_t follows

$$d_t = \rho d_{t-1} + \varepsilon_t, \ |\rho| < 1 \tag{2}$$

where ε_t is an $iid(0, \sigma^2)$ random variable (such that $\sigma^2 < +\infty$) with compact support $[-a, a]$, $a > 0$. Evans and Honkapohja (1999, 2001) consider different economic environments that also give rise to such specifications.

The assumption that the exogenous process for d_t has compact support is not very restrictive and clearly highlights our result: while the stationary distribution of an exogenous driving process has thin tails, the stationary distribution of the related endogenous variable may have fat tails, a result also characterized as "thin tails in, thick tails out". Furthermore, the assumption of compact support for ε_t makes it easy to show that the autoregressive exogenous process is uniformly recurrent over its stationary distribution. The assumption of uniform recurrence simplifies proofs and is further discussed in detail in the next section.

Anticipating our empirical application, we briefly provide an asset pricing interpretation for the model in (1)-(2). Following Lucas (1978), a single asset endowment economy with utility over consumption given by

$$u(C_t) = \frac{C_t^{1-\gamma}}{1-\gamma}, \ \gamma > 0 \tag{3}$$

yields, under a no-bubbles condition, the nonlinear pricing equation

$$P_t = E_t \left\{ \beta \left(\frac{D_{t+1}}{D_t} \right)^{-\gamma} (P_{t+1} + D_{t+1}) \right\} \tag{4}$$

where $\beta \in (0, 1)$ is the usual exponential discount factor and (real) dividends (D_t) follow

some exogenous stochastic process. Log-linearizing the above equation yields

$$p_t = \beta E_t(p_{t+1}) + (1 - \beta - \gamma)E_t(d_{t+1}) + \gamma d_t \tag{5}$$

where all lowercase variables denote log-deviations from the steady state $(\overline{P}, \overline{D}) = \left(\frac{\delta}{1-\delta}, 1\right)$. The exogenous process for d_t follows the same specification as above and since $E_t(d_{t+1}) = \rho d_t$,

$$p_t = \beta E_t(p_{t+1}) + \theta d_t,\ \theta \equiv (1 - \beta - \gamma)\rho + \gamma \tag{6}$$

is the fundamental expectational difference equation for prices.[6]

Returning to our linear model of learning, we follow Evans and Honkapohja (1999, 2001) and assume that the perceived law of motion (PLM) of the representative agent is

$$p_t = \phi_{t-1} d_{t-1} + \xi_t,\ \xi_t \sim i.i.d.(0, \sigma_\xi^2),\ \sigma_\xi^2 < +\infty, \tag{8}$$

which in turn implies

$$E_t(p_{t+1}) = \phi_{t-1} d_t, \tag{9}$$

where ϕ_{t-1} is the coefficient that agents estimate from the data to forecast p_t. Inserting the

[6] The rational expectations solution to (6) is

$$p_t = \phi^{REE} d_t,\ \phi^{REE} = \frac{\theta}{1 - \beta\rho} \tag{7}$$

for all $\beta\rho \neq 1$.

above into (6) yields the actual law of motion (ALM) under learning:[7]

$$p_t = \beta\phi_{t-1}d_t + \theta d_t = (\beta\phi_{t-1} + \theta)d_t \tag{10}$$

$$= (\beta\phi_{t-1} + \theta)\rho d_{t-1} + (\beta\phi_{t-1} + \theta)\varepsilon_t. \tag{11}$$

In contrast the ALM under rational expectations is

$$p_t = \phi d_t = \phi\rho d_{t-1} + \phi\varepsilon_t. \tag{12}$$

Under SGCG learning, ϕ_t evolves as[8]

$$\phi_t = \phi_{t-1} + g d_{t-1}(p_t - \phi_{t-1}d_{t-1}),\ g \in (0,1). \tag{13}$$

At this point we take the gain parameter g as given, but in section 5. we will estimate its value under our learning model with Bayesian agents who expect a random walk drift in ϕ. Following the usual practice in the literature for analyzing learning asymptotics, we insert

[7] We note that in the asset pricing context, the ALM is linear in the 'belief' parameter (ϕ_t). In other contexts the ALM might be nonlinear in beliefs. However, the linear forces generating large deviations in the adaptive learning model may drive the dynamics in nonlinear contexts. For example in Cho et al. (2002) adaptive learning leads to non-neglible probablities for large deviations even in the prescence of nonlinearities for the true data generating process.

[8] See Carceles-Poveda and Giannitsarou (2007, 2008) for details and derivations under a variety of learning algorithms.

the ALM under learning in place of p_t in the recursion for ϕ_t in (13) to obtain

$$\phi_t = \lambda_t \phi_{t-1} + \psi_t \tag{14}$$

$$\lambda_t = 1 - (1 - \rho\beta) g d_{t-1}^2 + \beta g d_{t-1} \varepsilon_t = 1 - g d_{t-1}^2 + g\beta d_t d_{t-1} \tag{15}$$

$$\psi_t = \theta\rho g d_{t-1}^2 + \theta g d_{t-1} \varepsilon_t = \theta g d_t d_{t-1}. \tag{16}$$

The equation in (14) takes the form of a linear recursion with both multiplicative (λ_t in (15)) and additive (ψ_t in (16)) noise. We show in the next Section that the stationary distribution of ϕ_t can be fat-tailed and indeed follows a power law even though the forcing variable (d_t) is a thin tailed process. Under the asset pricing application this implies that the price-dividend ratio (ϕ_t) can exhibit large deviations from its rational expectations equilibrium value with non-negligible probabilities.

3. Large Deviations and Rare Events

As noted, λ_t is a random variable generating multiplicative noise, and our main result is that it can be the source of large deviations and fat tails for the stationary distribution of ϕ_t. There are two elements that are absolutely critical for this result. First, the distribution of the random variable λ must have $E|\lambda| < 1$ or a stationary distribution fails to exist (see Brandt (1986)). Second, for ϕ_t to have a fat tail even if the exogenous driving process, the dividends, are thin tailed, we need the distribution of λ to have some support above the unit circle: $P(|\lambda| > 1) > 0$. Since the distribution of λ_t is governed by the exogenous process for d_t we will need some restrictions on $\{d_t\}_{t\in\mathbb{N}}$ as discussed below. In particular in section 5.

where we will apply our results to the asset pricing model and characterize the price-dividend ratio, these restrictions will apply to the stationary distribution of dividends.

We use results from large deviation theory (see Hollander (2000)) together with the work of Saporta (2005), Roitershtein (2007) and Collamore (2009) to characterize the tail of the distribution of ϕ_t.[9] Let $\mathbb{N} = 0, 1, 2...,$ and note that the stationary $AR(1)$ Markov process $\{d_t\}_{t\in\mathbb{N}}$ given by (2) is uniformly recurrent, and has compact support $\mathcal{D} = \left[\frac{-a}{1-\rho}, \frac{a}{1-\rho}\right]$ (see Nummelin (1984), p. 93).[10] We use the uniform recurrence of $\{d_t\}_{t\in\mathbb{N}}$ in step (ii) of the proof of 1 below to show that $|\lambda| > 1$ with positive probability, or $P_\omega(|\lambda| > 1) > 0$, which is essential to obtain fat tails for $\{\phi_t\}$.[11]

Next we seek restrictions on the support of the iid noise $\varepsilon_t \in [-a, a]$ to ensure that $E|\lambda_\infty| < 1$ where, from equation (15), λ_∞ is the random variable associated with the stationary distribution of d_t. For simplicity, in order to derive restrictions on a that assures $E|\lambda_\infty| < 1$ we assume that ε_t is uniformly distributed. We could just as easily have assumed another distribution, for example a triangular distribution, or even another skewed distribution over $[-a, a]$, and sought restrictions on its support, or a, to ensure that $E|\lambda_\infty| < 1$. The uniform distribution leads to easy computations, and makes it quite clear that it is not the skewness or the tails of the distribution of ε_t that drive our results on the tails of distribution

[9] For an application of these techniques to the distribution of wealth see Benhabib et al. (2011) and to regime switching, Benhabib (2010).

[10] To define uniformly recurrent let $(\mathbb{X}, \mathcal{X})$ be a measurable space and define ${}_BP^m(x, A) = \mathbb{P}(X_n \in A, X_i \notin B, m = 1, ...m-1)$. A chain $\{X_n\}$ is uniformly φ-recurrent if for all $A \in \mathcal{X}$ with $\varphi(A) > 0$, if $\lim_n \Sigma_{m=1}^n {}_AP^m(x, A) = 1$ holds uniformly in x. That is, for all $\varepsilon > 0$ there exists N such that for all $x \in \mathbb{X}$ and $n \geq N$, $\Sigma_{m=1}^n {}_AP^m(x, A) = 1 - \varepsilon$ (see Petritis (2012, Chapter 11)). To assure that the $AR(1)$ process $\{d_t\}_{t\in Z}$ is uniformly recurrent we also assume that the distribution of ε_t is not a singular (see Nummelin (1984, p. 92)). This is a very weak requirement: a probability distribution is singular on R^n if it is concentrated on a set of Lebesgue measure zero *and* gives probability zero to every one-point set. An example on R^1 would be the Cantor distribution, a probability distribution over a Cantor set.

[11] This requirement of uniform recurrence can be weakened, as discussed in Collamore (2009) in more detail, but proofs would become more cumbersome.

of the price dividend ratio. However a restriction on a that assures $E|\lambda_\infty| < 1$, no matter what the underlying distribution, is critical. If $E|\lambda_\infty| \geq 1$, then ϕ_t does not even have a limiting stationary distribution, so our results about fat tails cannot hold.

We assume for simplicity therefore that $\varepsilon_t \in [-a, a]$ and is uniform, and that[12]

$$a < \left(\frac{6\left(1-\rho^2\right)}{g\left(1-\beta\rho\right)} \right)^{0.5}. \tag{17}$$

Note that

$$\begin{aligned} E(\lambda_t) &= E\left(1 - g\left(d_{t-1}\right)^2 + g\beta\left(d_{t-1}\left(\rho d_{t-1} + \varepsilon_t\right)\right)\right) \\ E(\lambda_t) &= 1 - gE\left(d_{t-1}\right)^2 + \beta\rho g E\left(d_{t-1}\right)^2 \\ E(\lambda_\infty) &= \left(1 - gE\left(d_{t-1}\right)^2\left(1-\beta\rho\right)\right)_{t\to\infty}. \end{aligned}$$

Since ε_t is iid and is uniform with variance σ^2,

$$E(\lambda_\infty) = 1 - g\frac{\sigma^2}{1-\rho^2}\left(1-\beta\rho\right) \tag{18}$$

$$E(\lambda_\infty) = 1 - g\frac{\frac{1}{12}\left(2a\right)^2}{1-\rho^2}\left(1-\beta\rho\right). \tag{19}$$

From equation (19) it follows that $E(\lambda_\infty) < 1$, and solving for a such that $E(\lambda_\infty) > -1$, we obtain the restriction (17) to guarantee that $E|\lambda_\infty| < 1$, which is the only reason that we

[12] We can express this condition as

$$g < \frac{6(1-\rho^2)}{a^2(1-\beta\rho)}$$

which implies that given a, if g is too high, the condition $E|\lambda_\infty| < 1$ may fail and the dynamics of ϕ_t may explode. We thank a referee for pointing this out.

impose the restriction on a.

We denote the stationary distribution of $\{d_t\}_{t\in\mathbb{N}}$ by π. Since $\{d_t\}_{t\in\mathbb{N}} \in \mathcal{D}$ and $[\varepsilon_t]_{t\in\mathbb{N}} \in [-a, a]$ are bounded, so are $\{\lambda_t\}_{t\in\mathbb{N}}$ and $\{\psi_t\}_{t\in\mathbb{N}}$, and we define $(\lambda_t, \psi_t)_{t\in\mathbb{N}} \in \mathcal{B}$. In fact, following the definition of Roitershtein (2007), $\{d_t, (\lambda_t, \psi_t)\}_{t\in\mathbb{N}}$ constitutes a Markov Modulated Process (MMP) defined on the product space $(\mathcal{D}, \mathcal{B})$: conditional on d_t, the evolution of the random variables $\lambda_{t+1}(d_t, d_{t-1})$ and $\psi_{t+1}(d_t, d_{t-1})$ are given by

$$P(d_t \in A, (\lambda_t, \psi_t) \in B) = \int_A K(d, \mathfrak{d}y)\, G(d, y, B)\,|_{d=d_{t-1}}, \tag{20}$$

$$G(d, y, \cdot) = P((\lambda_t, \psi_t) \in \cdot) \mid d_{t-1} = d,\ d_t = y), \tag{21}$$

where $A \in \mathcal{D}$, $B \in \mathcal{B}$, $K(d, \mathfrak{d}y)$ is the transition kernel of the Markov process $\{d_t\}_{t\in\mathbb{N}}$ and $\mathfrak{d}y$ represents the differential. In other words an MMP does not require λ_t and ψ_t to be fully independent but allows a form of dependence where both can be driven by the process for $\{d_t\}_{t\in\mathbb{N}}$. In addition, since either or both can also be subject to *iid* shocks, they do not have to be perfectly correlated. Thus the probability that d_t will belong to a set A and (λ_t, ψ_t) will belong to a set B depends on d_{t-1} and on the the transition kernel of the Markov process $\{d_t\}_{t\in\mathbb{N}}$. This will in fact be the case when we apply our results to asset prices in section 5. where dividends drive both the multiplicative and the additive parts of the process for ϕ_t.

To set the stage for Proposition 1 let $S_n = \sum_{t=1}^{n} \log|\lambda_t|$. Following Roitershtein (2007)

and Collamore (2009)[13] the tail of the stationary distribution of $\{\phi_t\}_t$ depends on the limit[14]

$$\Lambda(\delta) = \lim_{n\to\infty} \sup \frac{1}{n} \log E \prod_{t=1}^{n} |\lambda_t|^{\delta} = \lim_{n\to\infty} \sup \frac{1}{n} \log E[\exp(\delta S_n)] \ \forall \ \delta \in \mathbb{R}. \tag{22}$$

Using results in Roitershtein (2007), we can now prove the following about the tails of the stationary distribution of $\{\phi_t\}_{t\in\mathbb{N}}$:

Proposition 1 *For π-almost every $d_0 \in [-a, a]$, there is a unique positive $\kappa < \infty$ that solves $\Lambda(\delta) = 0$, such that*

$$K_1(d_0) = \lim_{\tau\to\infty} \ \tau^{\kappa} P(\phi > \tau | d_0) \text{ and } K_{-1}(d_0) = \lim_{\tau\to\infty} \ \tau^{\kappa} P(\phi < -\tau | d_0) \tag{23}$$

and $K_1(d_0)$ and $K_{-1}(d_0)$ are not both zero.[15]

[13] For results on processes driven by finite state Markov chains see Saporta (2005).

[14] $\lim_{n\to\infty} \sup \frac{1}{n} \log E[\exp(\delta S_n)]$ is the Gartner Ellis limit that also appears in large deviations theory. For an exposition see Hollander (2000).

[15] *We can also show that $\pi(K_1(d_0) = K_{-1}(d_0)) = 1$ if a is large enough. This follows from Condition G given by Roitershtein (2007): Conditon G holds if there does not exist a partition of the irreducible set $D = \left\{ d \subset \left(\frac{-a}{1-\rho}, \frac{a}{1-\rho} \right) \right\}$ into two disjoint sets D_{-1} and D_1 such that:*

$$\begin{aligned} P(d &\in D_{-1}, \rho d + \varepsilon \in D_1, \lambda < 0) \\ &= P(d \in D_{-1}, \rho d + \varepsilon \in D_{-1}, \lambda > 0) = 0 \end{aligned}$$

where $\varepsilon \subset [-a, a]$ and $\rho \subset (0, 1)$. (See Roitershtein's Definition 1.7 and subsequent discussion, and his Proposition 4.1.) Suppose in fact that $P(d \in D_{-1}, \rho d + \varepsilon \in D_1, \lambda > 0) = 0$ for D_{-1} with minimal element d_0 and maximal element d_1. Then $P(d \in D_{-1}, \rho d + \varepsilon \in D_{-1}, \lambda > 0) = 1$. Then it must be true, since d_1 is the maximum element of D_{-1}, that $\rho d_1 + a \leq d_1$ and so $\frac{a}{1-\rho} \leq d_1$, implying $d_1 = \frac{a}{1-\rho}$. Similarly, it must be true that $\rho d_0 - a \geq d_0$ so that $\frac{-a}{1-\rho} \geq d_0$, implying $\frac{-a}{1-\rho} \geq d_0$. Thus $D_{-1} = D$, that is the whole set. Now we can show that for a large enough, $P(d \in D, \rho d + \varepsilon \in D, \lambda > 0) = 1$ cannot hold. Since

$$\lambda = 1 - g(d_0)^2 + g\beta d_0(\rho d_0 + \varepsilon) = 1 - g\left(d_0^2\right)(1 - \rho\beta) + g\beta d_0 \varepsilon,$$

we attain the smallest possible λ if we set $d_0 = \frac{a}{1-\rho}$ and $\varepsilon = -a$, or equivalently $d_0 = \frac{-a}{1-\rho}$ and $\varepsilon = a$. Then $\lambda \geq 0$ with probability 1 if and only if $a \leq \bar{a} = \frac{(1-\rho)}{(g(1+\beta(1-2\rho)))^{0.5}}$. If $a > \bar{a}$ with positive probability, then $P(\lambda < 0) > 0$, which contradicts $P(d \in D_{-1}, \rho d + \varepsilon \in D_{-1}, \lambda > 0) = 1$. Note also that $\lambda = 1$ for $d_0 = 0$ so it also follows that the $P(\lambda > 0) > 0$.

Proof. The results follow directly from Roitershtein (2007), Theorem 1.6 if we show the following:

(i) There exists a δ_0 such that $\Lambda(\delta_0) < 0$. First we note that $\Lambda(0) = 0$ for all n. Note also that

$$\begin{aligned}
\Lambda'(0) &= \lim_{n\to\infty} \sup \frac{1}{n} \frac{d \log E \prod_{t=1}^{n} |\lambda_t|^{\delta}}{d\delta} |_{\delta=0} \\
&= \lim_{n\to\infty} \sup \frac{1}{n} \left(E \prod_{t=1}^{n} |\lambda_t|^{\delta} \right)^{-1} E \left(\prod_{t=1}^{n} |\lambda_t|^{\delta} \log \prod_{t=1}^{n} |\lambda_t| \right) |_{\delta=0} \\
&= \lim_{n\to\infty} \sup \frac{1}{n} E \log \prod_{t=1}^{n} |\lambda_t|
\end{aligned}$$

For large n, as $\{\lambda_t\}_t$ converges to its stationary distribution ω, we have

$$\Lambda'(0) = \lim_{n\to\infty} \sup \frac{1}{n} \log E \prod_{t=1}^{n} |\lambda_t| = E_{\omega} \left(\log |\lambda_{\infty}| \right)$$

From equations (17)-(19) we have $E_{\omega} |\lambda_{\infty}| < 1$. Therefore $\Lambda'(0) = E_{\omega} \log (|\lambda_{\infty}|) < 0$, and there exists $\delta_0 > 0$ such that $\Lambda(\delta_0) < 0$.

(ii) There exists a δ_1 such that $\Lambda(\delta_1) > 0$. As in (i) above, we can evaluate, using Jensen's inequality,

$$\begin{aligned}
\Lambda(\delta) &= \lim_{n\to\infty} \sup \frac{1}{n} \log E \prod_{t=1}^{n} |\lambda_t|^{\delta} = \lim_{n\to\infty} \sup \frac{1}{n} \log E[\exp(\delta S_n)] \qquad (24) \\
&= \lim_{n\to\infty} \sup \log \left(E[\exp(\delta S_n)] \right)^{\frac{1}{n}} \geq \lim_{n\to\infty} \sup \log \left(E[\exp(\delta \frac{S_n}{n})] \right) \qquad (25)
\end{aligned}$$

so that at the stationary distribution of $\{\lambda_t\}_{t\in\mathbb{N}}$

$$\Lambda(\delta) \geq \log E_\omega[\exp(\delta \log|\lambda_\infty|)] = \log \int_\lambda [\exp(\delta \log|\lambda_\infty|)] d\omega(\lambda). \tag{26}$$

As $\delta \to \infty$ for $\log|\lambda| < 0$ we have $[\exp(\delta \log|\lambda_t|)] \to 0$, but if $P_\omega(\log|\lambda| > 0) > 0$ at the stationary distribution of $\{\lambda_t\}_t$, then $\lim_{\delta\to\infty} \Lambda(\delta) = \log \int_\lambda [\exp(\delta \log|\lambda_t|)] d\omega(\lambda) \to \infty$. Therefore if we can show that $P_\omega(\log|\lambda_t| > 0) > 0$, it follows that there exists a δ_1 for which $\Lambda(\delta_1) > 0$. Since $\Lambda(\delta)$ is convex[16], it follows that there exists a unique κ for which $\Lambda(\kappa) = 0$. To show that $P_\omega(|\lambda| > 1) > 0$, define $A = \left\{d \in \left(0, \frac{\mu a\beta}{1-\rho\beta}\right)\right\}$, $\mu \in (0,1)$ so that $\frac{\mu a\beta}{1-\rho\beta} < \frac{a}{1-\rho}$. At its stationary distribution $\{d_t\}_{t\in\mathbb{N}}$ is uniformly recurrent over $\left[\frac{-a}{1-\rho}, \frac{a}{1-\rho}\right]$ which implies that $P_\pi(d_{t-1} \in A) > 0$. We have $\lambda_t = 1 - \beta g d_{t-1}\left(\beta^{-1}(1-\rho\beta)d_{t-1} - \varepsilon_t\right)$, so for $d_{t-1} \in A$ and $\varepsilon_t \in (\mu a, a]$, it follows that $\lambda_t > 1$. Thus $P_\omega(|\lambda_t| > 1) = P_\pi(d_{t-1} \in A_t) P(\varepsilon_t \in (\mu a, a]) > 0$.

(iii) The non-arithmeticity assumption required by Roitershtein (2007) (p. 574, (A7)) holds[17]: There does not exist an $\alpha > 0$ and a function $G : \mathcal{R} \times \{-1, 1\} \to R$ such that

$$P(\log|\lambda_t| \in G(d_{t-1}, \eta) - G(d_t, \eta \cdot sign(\lambda_t)) + \alpha\mathbb{N}) = 1. \tag{27}$$

We have

$$\log|\lambda_t| = \log\left|(1 - g d_{t-1}^2 + g\beta d_t d_{t-1})\right| = \log\left|\left(1 - (1-\rho\beta) g d_{t-1}^2 + \beta g d_{t-1}\varepsilon_t\right)\right| = F(d_{t-1}, \varepsilon_t), \tag{28}$$

[16] This follows since the moments of nonnegative random variables are log convex (in δ); see Loeve (1977, p. 158).

[17] See also Alsmeyer (1997). In other settings $\{\lambda_t\}_t$ may contain additional *iid* noise independent of the Markov Process $\{d_t\}_t$, in which case the non-aritmeticity is much more easily satisfied.

which contains the cross-partial term $d_t d_{t-1}$. Therefore in general $F(d_{t-1}, \varepsilon_t)$ cannot be represented in separable form as $R(d_{t-1}, \eta) - R(d_t, \eta) + \alpha\mathbb{N} \ \ \forall (d_{t-1}, d_t)$ where $d_t = \rho d_{t-1} + \varepsilon_t$. Suppose to the contrary that there is a small rectangle $[D, D^*] \times [E, E^*]$ in the space of (d, ε), over which λ remains of constant sign, say positive, such that $F(d, \varepsilon) = R(d) - R(\rho d + \varepsilon)$, d is in the interior of $[D, D^*]$, and ε is in the interior of $[E, E^*]$, up to a constant from the discrete set $\alpha\mathbb{N}$, which we can ignore for variations in $[D, D*] \times [E, E*]$ that are small enough. Now fix d, d' close to one another in the interior of $[D, D^*]$. We must have, for $\varepsilon \in [E + \rho|d - d'|, \ E^* - \rho|d - d'|]$, that

$$\begin{aligned} F(d, \varepsilon) - R(d) &= -R(\rho d + \varepsilon) = -R(\rho d' + \varepsilon + \rho(d - d')) \quad (29) \\ &= F(d', \varepsilon + \rho(d - d')) - R(d'), \quad (30) \end{aligned}$$

or $F(d, \varepsilon) - F(d', \varepsilon + \rho(d - d')) = R(d) - R(d')$. However the latter cannot hold since the cross-partial term $d_{t-1}\varepsilon_t$ in $F(d_{t-1}, \varepsilon_t) = 1 - (1 - \rho\beta) g d_{t-1}^2 + \beta g d_{t-1}\varepsilon_t$ is non-zero except for a set of zero measure where d or ε are zero.[18,19]

(iv) To show that $K_1(d_0) = \lim_{\tau\to\infty} \ \tau^\kappa P(\phi > \tau | d_0)$ and $K_{-1}(d_0) = \lim_{\tau\to\infty} \ \tau^\kappa P(\phi <$

[18] We thank Tomasz Sadzik for suggesting this proof for (iii).

[19] We can avoid possible degeneracies that may occur if λ_t and ψ_t have a specific form of dependence so that

$$P(\phi | \lambda_t \phi + \psi_t = \phi) = 1.$$

Note

$$\begin{aligned} \phi &= \frac{\psi_t}{1 - \lambda_t} = \frac{\theta\rho g d_t^2 + \theta g d_t \varepsilon_{t+1}}{1 - (1 - \rho\beta) g d_t^2 + \beta g d_t \varepsilon_{t+1}} \\ &= \frac{\theta}{\beta} \frac{\beta\rho g d_t^2 + g\beta g d_t \varepsilon_{t+1}}{1 - (1 - \rho\beta) g d_t^2 + \beta g d_t \varepsilon_{t+1}} \end{aligned}$$

Differentiating with respect to ε_t, the right side is zero only if $\beta\rho g d_t^2 = 1 - (1 - \rho\beta) g d_t^2$, or $\beta\rho g = 1 - g + g\rho\beta$. This holds only if $g = 1$. So in general, for any d_t, there exists a constant ϕ such that $P(\phi | \lambda_t \phi + \psi_t = \phi) = 1$ only if $g = 1$, which we ruled out by assumption.

$-\tau|d_0)$ are not both zero, we have to assure, since ψ_t and λ_t are not assumed to be independent, that ϕ is not a deterministic function of the initial d_{-1}. We invoke (a) and (c) of Proposition 8.1 in Roitershtein (2007): Condition 1.6, $\pi\left(K_1\left(d_0\right)+K_{-1}\left(d_0\right)=0\right)=1$, holds if and only if there exists there exists a measurable function Γ: $\left[\frac{-a}{1-\rho},\frac{a}{1-\rho}\right]\rightarrow R$ such that

$$P\left(\psi_0+\lambda_0\Gamma\left(\rho d_{-1}+\varepsilon_0\right)=\Gamma\left(d_{-1}\right)\right)=1.$$

However

$$\psi_0+\lambda_0\Gamma\left(\rho d_{-1}+\varepsilon_0\right)=\theta g d_{-1}\rho d_{-1}+\theta g d_{-1}\varepsilon_0+\left(1-g d_{-1}^2+g\beta d_{-1}\left(\rho d_{-1}+\varepsilon_0\right)\right)\Gamma\left(\rho d_{-1}+\varepsilon_0\right)$$

is a random variable that depends on ε_0 while $\Gamma\left(d_{-1}\right)$ is a constant, so

$$P\left(\psi_0+\lambda_0\Gamma\left(\rho d_{-1}+\varepsilon_0\right)=\Gamma\left(d_{-1}\right)\right)<1$$

and Condition 1.6 in Roitershtein (2007) cannot hold. Then from Roitershtein (2007) Proposition 1.8 (c), $K_1\left(d_0\right)$ and $K_{-1}\left(d_0\right)$ are not both zero.[20] ■

The Proposition above characterizes the tail of the stationary distribution of ϕ as a power tail with exponent κ. It follows that the distribution of ϕ has moments only up to the highest integer less than κ, and is a 'fat tailed' distribution rather than a Normal. The results are driven by the fact that the stationary distribution of $\{\lambda_t\}_{t\in\mathbb{N}}$ has a mean less

[20] In models where the driving stochastic process is *iid* or is a finite stationary Markov chain, the exponent κ can be analytically derived using the results of Kesten (1973) and Saporta (2005). In the case where λ is *iid* in equation (14), κ solves $E\left(\lambda^{\kappa}\right)=1$. In the finite Markov chain case, under appropriate assumptions, κ solves $\varsigma\left(PA^{\kappa}\right)=1$ where P is the transition matrix, A is a diagonal matrix of the states of the Markov chain assumed to be non-negative, and $\varsigma\left(PA^{\kappa}\right)$ is the dominant root of PA^{κ}.

than one, which tends to induce a contraction towards zero, but also has support above 1 with positive probability, which tends to generate divergence towards infinity. The stationary distribution arises out of a balance between these two forces. Then large deviations as strings of realizations of λ_t above one, even though they may be rare events, can produce fat tails.

In the asset price model ϕ relates the dividends to asset prices. Under adaptive learning, the results above show how the probability distribution of large deviations, or 'escapes' of ϕ from its REE value is characterized by a fat tailed distribution, and will occur with higher likelihood than under a Normal.[21]

We now briefly discuss the case where $\{d_t\}_t$ is an $MA(1)$ process. Proposition 1 still applies and we obtain similar results to the $AR(1)$ case. Let

$$d_t = \varepsilon_t + \zeta\varepsilon_{t-1},\ |\zeta| < 1,\ t = 1, 2... \tag{31}$$

Then at its stationary distribution $d_t \in [-a\,(1+\zeta)\,, a\,(1+\zeta)]$. Under the PLM

$$p_t = \phi_{0t}\varepsilon_t + \phi_{1t}\varepsilon_{t-1}, \tag{32}$$

after observing ε_t at time t but not ϕ_{1t+1}, the agents expect

$$E_t(p_{t+1}) = \phi_{0t}E_t(\varepsilon_{t+1}) + \phi_{1t}E_t(\varepsilon_t) = \phi_{1t}\varepsilon_t. \tag{33}$$

[21] In the model of Cho et al. (2002), the monetary authority has a misspecified Philips curve and sets inflation policy to optimize a quadratic target. The learning algorithm using a constant gain however is not linear in the recursively estimated parameters (the natural rate and the slope of the Philips curve).

Then the ALM is

$$p_t = \beta\phi_{1t}\varepsilon_t + \gamma(\varepsilon_t + \zeta\varepsilon_{t-1}) = [\beta\phi_{1t} + \gamma]\,\varepsilon_t + \gamma\zeta\varepsilon_{t-1}$$

and the REE is given by

$$\phi_0 = \gamma(1 + \beta\zeta), \tag{34}$$

$$\phi_1 = \gamma\zeta. \tag{35}$$

Under the learning algorithm in equation (13) we obtain

$$\phi_{1t} = \phi_{1t-1} + gd_{t-1}(p_t - \phi_{1t-1}d_{t-1}), \tag{36}$$

$$\phi_{1t+1} = \lambda_{t+1}\phi_{1t} + \psi_{t+1}, \tag{37}$$

$$\lambda_{t+1} = 1 - gd_t^2 + g\beta\varepsilon_{t+1}d_t, \tag{38}$$

$$\psi_{t+1} = g\gamma\varepsilon_{t+1}d_t + \gamma\zeta g d_t\varepsilon_t. \tag{39}$$

It is straightforward to show that at the stationary distribution of $\{\lambda_t\}_t$, $E(\lambda_t) < 1$, and that $P(\lambda_t > 1) > 0$. It is also easy to check that $\lambda_t > 0$ if $a < ((1+\zeta)(1+\zeta-\beta))^{-0.5}$. With the latter restriction, it is easy to check that the other conditions in the proof of Proposition 1 are satisfied.

4. Model Simulations and Comparative Statics

The theoretical results above indicate that, in the context of a simple asset pricing model, rare but large shocks to the exogenous dividend process can throw off forecasts for the price-dividend ratio away from its rational expectation value. Of course escapes are more likely if the variance of the shocks to dividends are high. More critically, escapes in the long-run are possible if agents put a large weight on recent observations and discount older ones. The decay of the weights on past observations depends on the gain parameter g.[22] The size of the Bayesian optimal g will in turn depend on the drift that agents expect in the estimated parameter ϕ. We will estimate g in the next section, both directly, and also from the perspective of Bayesian agents expecting a random walk drift in ϕ.

In this section we explore how κ is related to the underlying parameters of our model. We can simulate the learning algorithm that updates ϕ, and then estimate κ from the simulated data using a maximum likelihood procedure following Clauset et al. (2009). We can then explore how κ varies as we vary model parameters. We simulate 1000 series, each of length 5000, for ϕ_t under the $AR(1)$ assumption for dividends with iid uniform shocks. We then feed the simulated series into the model to produce $\{P_t\}$ and $\{P_t/D_t\}$. We estimate κ for each simulation and produce an average κ.

Escapes or large deviations in prices will take place when sequences of consistently large shocks to dividends (in absolute value) throw off the learning process away from the rational expectations equilibrium. Such escapes will be more likely if dividend shocks can produce

[22] Under constant gains the decay in weights on past observations dating i periods back is given by $(1-g)^{i-1}$. Note of course that the value of g computed with annual data would be larger than the corresponding g if the data were converted to quarterly.

values of λ_t above 1, as we can see from equations (14-16). We expect lower κ, or fatter tails, as the support of λ_t that lies above 1 gets larger.

In the $AR(1)$ case for dividends we have $\lambda_{t+1} = 1 - (1 - \rho\beta)gd_t^2 + \beta g d_t \varepsilon_{t+1}$. Given the stationary distribution of $\{d_t\}_t$ and that of $\{\varepsilon_t\}_t$, the support of λ_t above 1 unambiguously increases if β increases. In principle increasing ρ can have an ambiguous effect: while the term $(1-\beta\rho)$ declines and tends to raise λ_t for realizations of d_t and ε_{t+1}, the support of the stationary distribution of $\{d_t\}_t$ gets bigger with higher ρ. While this can increase $(1-\rho\beta)gd_t^2$ and reduce the support of λ that is above 1 for large realizations of d_t^2, in our simulations the former effect seems to dominate. Finally we expect that decreasing g will shrink the support of λ_t that is above 1 so that κ increases with g: as the gain parameter decreases, the tails of the stationary distribution of $\{\phi_t\}$ get thinner.[23]

We use a baseline parameterization, $(\rho, g, \beta, \gamma) = (0.80, 0.4, 0.95, 2.5)$ based on estimates that we obtain in the next section. The estimated parameters, except for g, are in line with standard calibrations. The discount factor of $\beta = 0.95$ is consistent with annual data and an annual discount rate of about 5%. While empirical estimates of g are hard to come by, the usual values of g used in theoretical models are much smaller, in the order of 0.01 or 0.04, suggesting a very slow decay in the weights attached to past observations. Values of g in the range of 0.3-0.5 indicate a high decay rate, suggesting a propensity for the agents to think that "this time it's different". As noted above, we attempt to estimate g in the context of

[23] This of course is in accord with the Theorem 7.9 in Evans and Honkapohja (2001). As the gain parameter $g \to 0$ and $tg \to \infty$, $\{\phi_t^g - \varkappa\}/g^{0.5}$ converges to a Gaussian variable where $\varkappa$ is the globally stable point of the associated ODE describing the mean dynamics. More generally, as $g \to 0$, the estimated coefficient under learning with gain parameter g, ϕ_t^g, converges in probability (but not uniformly) to $\varkappa$ for $t \to \infty$. However, there will always exist arbitrarily large values of t with ϕ_t^g taking values remote from $\varkappa$ (see Benveniste et al. (1980), pp. 42-45). Note however that our characterization of the tail of the stationary distribution of $\{\phi_t\}_t$ and of κ is obtained for fixed $g > 0$.

our model by two separate methods in the next section. However, as the comparative statics in Figure 2 below demonstrate, for the learning model to explain the fat tails and the high variance of the P/D ratio, the gain parameter has to be large enough. This also implies, as discussed further in the next section, that the expected drift in the estimated parameters should have a large variance.

For a parametrization based on the asset pricing model, we set the value of $a = 0.33$ to match the standard deviation of linearly detrended dividends in the data. We find that the average κ is 5.0210, the average price-dividend ratio (P_t/D_t) is 20.6274 and the average standard deviation of (P_t/D_t) is 9.8934. We then vary each element of $(\rho, g, \beta, \gamma, \alpha)$ while keeping the others at their baseline values. The results of varying each parameter around the baseline values are plotted in Figures 1 and 2 below.[24]

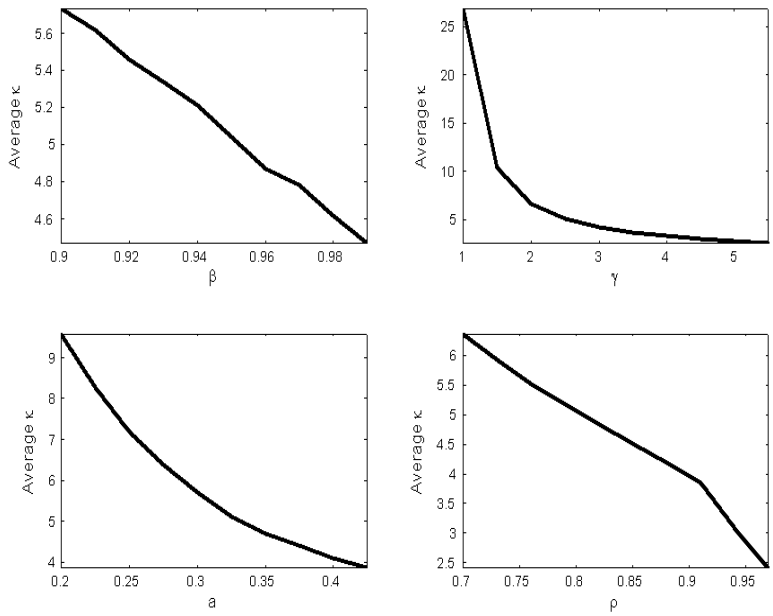

Figure 1. Simulation Results.

[24]The restriction given by equation (17) implies a maximum value of $a = \hat{a} = 4.2733$, the corresponding value for quarterly data would be 3.9933 (see the Quarterly Frequency Results Appendix). For all parameter values used to produce Figures 1 and 2, the restriction is easily satisfied.

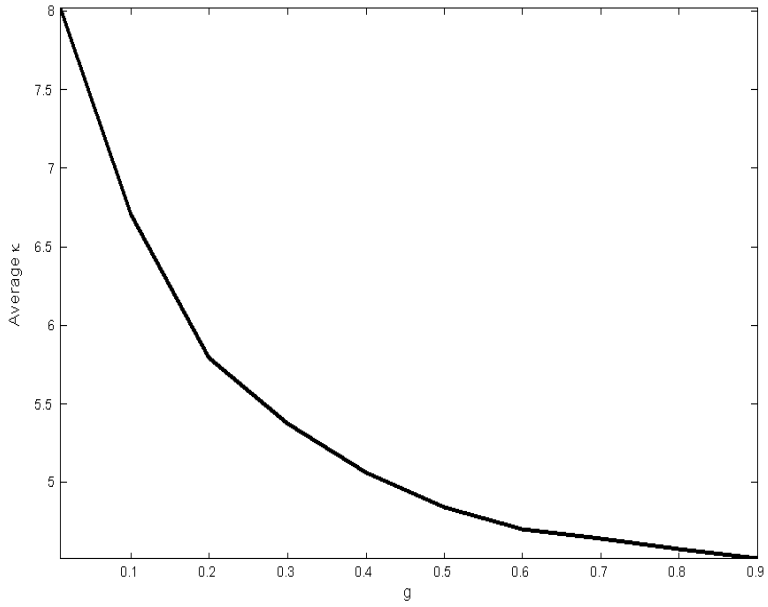

Figure 2. Simulation Results (cont'd.).

The simulation results confirm the notion that the average κ's should decline with β, γ and a. Figure 2 plots the results of the critical learning parameter g; it clearly demonstrates that as the learning gain falls, that is, the horizon for learning increases, the average κ rises. In summary, SGCG learning leads to large deviations of (P_t/D_t) from its rational expectations value even though the exogenous driving process for dividends is thin-tailed.

5. An Empirical Application

Figures 3-4 plot aggregate annual stock prices and dividends in the U.S. as measured by the S&P 500 and CRSP datasets. The plots show that, as predicted by standard theory, prices and dividends do move in tandem. However the price-dividend ratio, shown in the third panel of each Figure, exhibits large fluctuations, especially in the latter parts of the

sample.[25] These large fluctuations in the price-dividend ratio are difficult to explain with the standard rational expectations asset pricing model, for example that of Lucas (1978).[26]

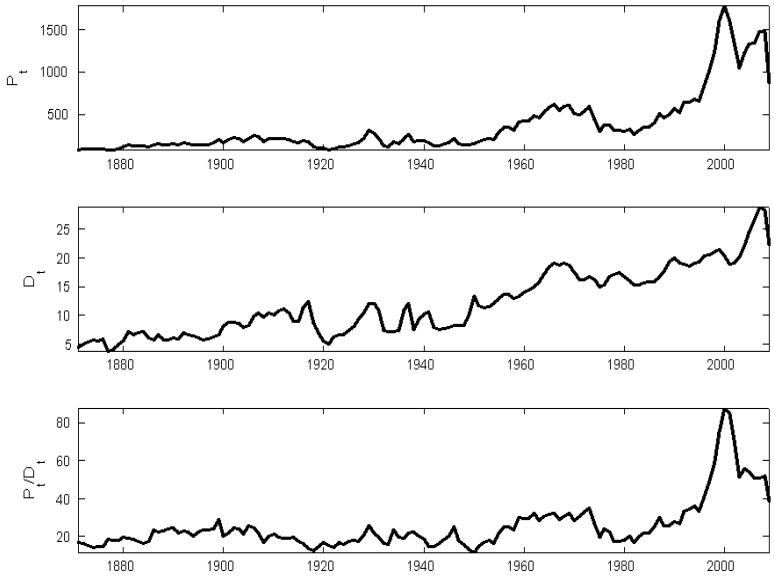

Figure 3. Annual S & P 500 (1871-2010).

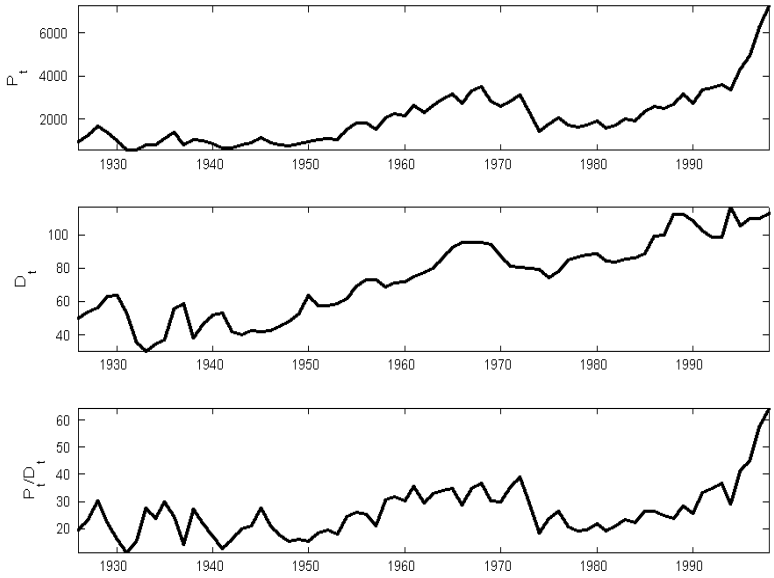

Figure 4. Annual CRSP (1926-1998).

We first check whether real world data on price-dividend ratios have fat tails. We use the maximum likelihood procedure following Clauset et al. (2009) to estimate κ associ-

[25] Details on the data employed are presented in the Data Appendix.

[26] See for example Carceles-Poveda and Giannitsarou (2008).

ated with P_t/D_t for both S&P 500 and CRSP dividend series plotted in Figures 3 and 4 above. The results provided in Table 1 below show fairly small values of κ for both series, suggesting that only the first few moments of P_t/D_t exist irrespective of the data source. Table 1 also reports the estimated persistence ρ under an $AR(1)$ specification for the two linearly detrended dividends series, alongside the average price-dividends ratio (P_t/D_t) and its standard deviation.[27]

Table 1. Data Characteristics

	S & P 500	CRSP
	1871-2010	1926-1998
$\widehat{\kappa}$	3.6914	5.5214
$s.e.(\widehat{\kappa})$	0.3828	2.6046
$\widehat{\rho}$	0.7891	0.7519
$s.e.(\widehat{\rho})$	0.0523	0.0777
Mean (P_t/D_t)	25.5211	26.1805
Std. Dev. (P_t/D_t)	13.1758	9.3298
Corr (P_t/D_t)	0.9438	0.7872
$r = \frac{\overline{D}_t}{\overline{P}_t}$	0.0336	0.0360
$\beta = (1 + r/4)^{-1}$	0.9917	0.9911
σ_d	0.1892	0.1649

[27] Whenever we employ actual dividends series, we linearly detrend (see DeJong and Dave (2011)). Note also that the dividends data have a higher standard deviation than that which is obtained only with post WWII data. This is because our data series also capture the Great Depression, and in the case of the S&P 500, the higher volatility in stock prices subsequent to the U.S. Civil War.

We use two separate approaches to get estimates for the gain parameter g. First we feed the actual S&P and CRSP dividend series into our learning model and estimate the parameters, $\vartheta = [g\ \gamma\ \beta\ \rho]$ by minimizing the squared difference between the empirical κ's reported in Table 1 and those generated by our model. That is, we implement a simulated minimum distance method to estimate ϑ as[28]

$$\min_{\vartheta} [\kappa - \kappa(\vartheta)]^2. \tag{40}$$

This estimation process necessarily puts a great deal of emphasis on the tail of the empirical data given by κ. Since the puzzle lies in the fat tail and high variance of P/D, emphasizing the tail in the estimation method may be justified. The parameter estimates other than g are certainly in line with basic calibrations in the literature, but the value of g, as expected from our model, is higher than the usual values of 0.01-0.04 that we find in the literature.

The minimization procedure is as follows. For candidate parametrizations of ϑ we employ the S&P 500 or CRSP series dividends d_t to calculate ϕ_t as per (14)-(16). The ALM (10) then produces a corresponding p_t series which in turn delivers a price-dividend ratio P_t/D_t. We then estimate the κ associated with the 'simulated' P_t/D_t, using the methods of Clauset et al. (2009) to produce the $\kappa(\vartheta)$. The minimization procedure searches over the parameter space of ϑ to implement (40). Table 2 below reports the estimates and associated standard errors for each of the S&P 500 or CRSP dividend series. We also report associated κ values

[28] Minimization was conducted using a simplex method and standard errors were computed using a standard inverse Hessian method.

obtained by simulating prices using the estimated parameters and the actual dividend data.[29]

Table 2. Parameter Estimates

	S & P 500		CRSP	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
g	0.3468	2.7158	0.5257	0.4722
γ	2.6503	1.7481	2.4598	0.6259
β	0.9615	0.3870	0.8984	0.4576
ρ	0.8729	0.0552	0.7959	0.1355
Associated κ	2.4128		5.5214	

The point estimates of g, ranging from 0.35 to 0.53 are high, although the standard errors are quite large, especially in the case of the S & P 500 dataset. The high estimates for the gain parameter g, which imply a fat tail for the price dividend ratio, reflect the fat tail (or low κ) that we observe in the data in Table 1, as expected. Standard errors are significantly smaller when we construct and use longer quarterly data.[30] Carceles-Poveda and Giannitsarou (2008) discuss possible values of g. Looking at standard deviations of the price-dividend ratios for the Lucas asset pricing model, they report that the standard deviations generated by the rational expectations or the learning models are smaller than the standard deviations in the actual data by factors of about 20 to 50. Note that our estimates of the parameter values, including g, are very close to those used by Carceles-Poveda and Giannitsarou (2008) in their simulations except for γ, the CRRA parameter: they set $\gamma = 1$ while we have it at $\gamma = 2.5$. Note also that for our simulations in Figure 1 κ drops

[29] Starting values for the minimization procedure were $\vartheta_0 = [0.5\ 2.5\ 0.95\ 0.75]$.

[30] Table 5 in the Quarterly Frequency Results Appendix provides estimates for quarterly data.

dramatically with γ.

For our second approach to pin down the gain parameter we let the agent optimally determine g by estimating the standard deviations of the parameter drift, the noise in the P/D ratio, and the shock to the dividend process.[31] Recall that under SGCG learning ϕ_t evolves as

$$\phi_t = \phi_{t-1} + gd_{t-1}(p_t - \phi_{t-1}d_{t-1}),\ g \in (0,1) \tag{41}$$

Consider the case in which the agents assume that the PLM is

$$p_t = \phi_{t-1}d_{t-1} + \xi_t,\ \xi_t \sim iid(0, \sigma_\xi^2),\ \sigma_\xi^2 < +\infty \tag{42}$$

with the coefficient ϕ drifting according to a random walk:

$$\phi_t = \phi_{t-1} + \Lambda_t,\ \Lambda_t \sim iid(0, \sigma_\Lambda^2),\ \sigma_\Lambda^2 < +\infty. \tag{43}$$

In this case, the Bayesian agent would use (41) to estimate σ_Λ, σ_d and σ_ξ and set an optimal estimate of the gain in the limit as

$$g = \frac{\sigma_\Lambda \sigma_d}{\sigma_\xi} \tag{44}$$

where σ_d denotes the standard deviation of d_t (see Evans et al (2010)). Under this approach, the long-run value of g that generates $\{p\}$ and $\{\phi\}$ under adaptive learning would be self-confirming, in the sense that agents would in fact estimate g using (44).

[31] See Sargent et al. (2006) and others for a more complex version of this approach for models requiring dynamic tracking estimation.

To compute (44) an estimate of σ_d is of course readily obtained from the dividend data. However we need to specify a method for the agents to compute estimates of σ_Λ and σ_ξ. If we recognize the system above as being analogous to a time varying parameter formulation, then employing the methods laid out in Kim and Nelson (1999) we can obtain estimates of σ_Λ and σ_ξ.[32] We report these results in Table 3 below.

Table 3. Drifting Beliefs Model Parameter Estimates

	S & P 500		CRSP	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
σ_Λ	0.8122	0.7718	0.8588	0.2963
σ_ξ	0.3157	0.0230	0.2596	0.0291
$\log L$	-61.4102		-17.5256	
σ_d	0.1892		0.1649	
Associated g	0.4866		0.5455	

These estimates suggest values of the gain significantly larger than those usually assumed in the literature.[33] Looking at Figure 2, a value of $g = 0.4866$ yields a tail estimate κ of about 4.9 while a value of $g = 0.5455$ yields a κ of about 4.75, compared to κ in the data ranging from 3.7 to 5.5 in Table 1. We also simulated the model with baseline parameter values but with gains of 0.4866 and 0.5455. These simulations resulted in average price-dividend ratios of 20.6324 and 20.6965 respectively with corresponding standard deviation values of 10.0051 and 10.5870.

[32] Given our estimate of $\beta = 0.95$ we convert the CRSP data to annual, summing dividends quarterly dividends for each year. For the S&P 500 we use the annual data reported by Shiller (1999), pp.439-441.

[33] We also apply block bootstrap methods to this estimation, detailed at the end of the Quarterly Frequency Results appendix. These methods allow us to construct average associated g's with attendant standard deviations, instead of the associated g in the last row of Table 3.

Finally, instead of using actual P and D data series, we generate data by simulating our model with our benchmark values $(\rho, g, \beta, \gamma) = (0.80, 0.4, 0.95, 2.5)$, and then compute g from (44) using the methods in Kim and Nelson (1999) to check that we recover a value close to 0.4.[34] The average g is 0.3826, which is quite close to and confirms the benchmark value of $g = 0.4$ that we used in generating the simulated data. In fact we conducted these simulations for a range of values of the gain parameter and then computed the associated average g value. Our prior was that the resulting plot would intersect near a gain value which was near $g = 0.4$; we provide the plot in Figure 5 below and note that our prior was confirmed.

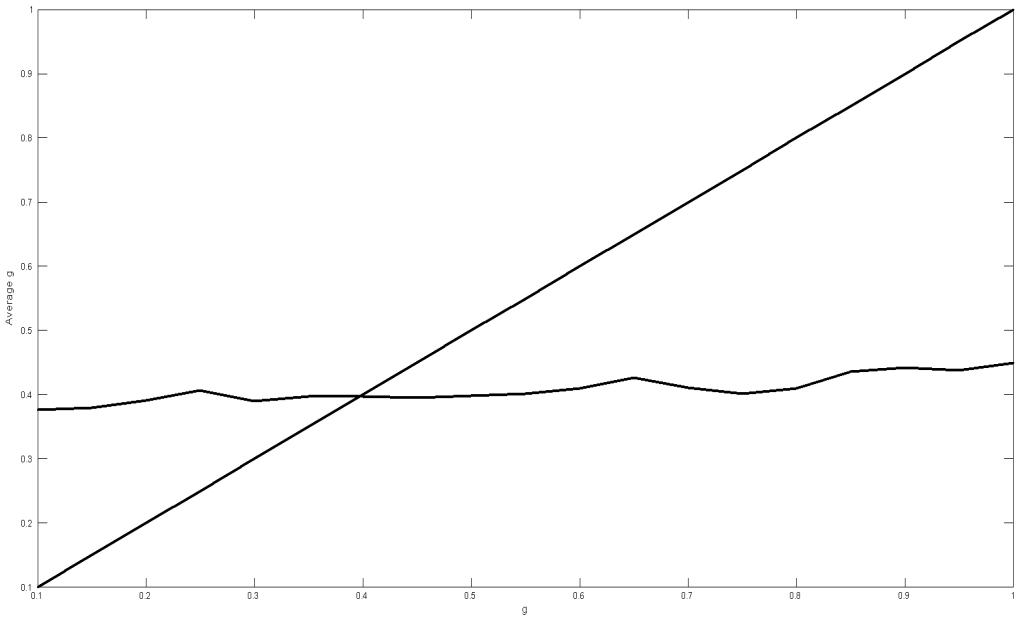

Figure 5. Fixed Point Plot.

[34] We run 1000 simulations each with 5000 periods, and obtain the average g from (44) across the 1000 simulations.

Given these results, we offer a caveat with respect to interpretation of the large gain parameter estimates that we find. One approach is to say that the model with SGCG learning does very well in matching features of the data given the gain parameter estimates. Alternatively one could argue that absent large estimates for the gain, the model does not do well in matching asset pricing facts. We prefer the former interpretation given that it is entirely reasonable to interpret our empirical exercise as suggesting that heavily discounting past observations is consistent with market participants' behavior. With annual data the estimates suggest market participants' horizon for learning is about 2-3 years which is not entirely unreasonable given the frequent swings in the data.

6. Conclusion

An important and growing literature replaces expectations in dynamic stochastic models not with realizations and unforecastable errors, but with regressions where agents 'learn' the rational expectations equilibria (REE). In these adaptive learning models when agents employ constant gain algorithms that put heavier emphasis on recent observations and shown to be optimal when there is drift in estimated parameters, escape dynamics can propel estimated coefficients away from the REE values. We show that in a constant gain adaptive learning model, the stationary distribution of the variables that agents are learning can be fat tailed, and that the tail index of this distribution can be characterized in terms of the parameters of the model.

We then analyze, in an asset pricing context, the stationary distribution of the price-dividend ratio in a canonical model with constant gain adaptive learning. We reinterpret

the learning algorithm as a linear recursion with multiplicative noise and use techniques from large deviations theory to characterize the tail of the stationary distribution of the price-dividend ratio.

In an asset pricing context 'bubbles', or asset price to dividend ratios that exhibit large deviations from their REE values (even though our model has presumed a no-bubble condition) can occur with a frequency associated with a fat tailed power law, as observed in the data. The techniques used in our paper can be generalized to higher dimensions, to finite state Markov chains, to continuous time,[35] and can be applied more generally to other economic models that use constant gain learning.

[35] See for example Saporta (2005), Saporta and Yao (2005), and Ghosh et al. (2010).

References

[1] Adam, K., Marcet, A., and J. P. Nicolini, 2008. "Stock Market Volatility and Learning," *European Central Bank Working Paper Series*, No. 862.

[2] Adam, K. and Marcet, A., 2011. "Internal Rationality, Imperfect Market Knowledge and Asset Prices," *Journal of Economic Theory*, 146, 1224-1256.

[3] Alsmeyer, G, 1997. "The Markov Renewal Theorem and Related Results," *Markov Process Related Fields* 3 103–127.

[4] Barro, R. J, 2009. "Rare Disasters, Asset Prices, and Welfare Costs," *American Economic Review*, 99:1, 243–264.

[5] Benhabib, J. Bisin, A. and S. Zhu, 2011. "The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents," *Econometrica* 79, 123-158.

[6] Benhabib, J., 2010. "A Note Regime Switching, Monetary Policy and Multiple Equilibria," *NBER Working Paper* No. 14770.

[7] Benveniste, A., Métivier, M. and P. Priouret, 1980. *Adaptive Algorithms and Stochastic Approximations,* Springer-Verlag, New York.

[8] Branch, W., and Evans, G. W., 2010, "Asset Return Dynamics and Learning," *Review of Financial Studies*, 1651-1680.

[9] Brandt, A., 1986. "The Stochastic Equation $Y_{n+1} = A_n Y_n + B_n$ With Stationary Coefficients," *Advances in Applied Probability*, 18, 211–220.

[10] Brennan, M. J., Xia, 2001. "Stock Price Volatility and Equity Premium," *Journal of Monetary Economics,* 47, 249-283.

[11] Bullard, J., Duffy, J., 2001. "Learning and Excess Volatility," *Macroeconomic Dynamics* 5, 272-302.

[12] Campbell, J. Y., 2003. "Consumption-Based Asset Pricing," *Handbook of the Economics of Finance*, George Constantinides, Milton Harris, and Rene Stulz eds., North-Holland, Amsterdam.

[13] Carceles-Poveda, E., Giannitsarou, C., 2007. "Adaptive Learning in Practice," *Journal of Economic Dynamics and Control* 31, 2659-2697.

[14] Carceles-Poveda, E., Giannitsarou, C., 2008. "Asset Pricing with Adaptive Learning," *Review of Economic Dynamics* 11 629–651.

[15] Chevillon, G., Mavroeidis, S., "Learning Generates Long Memory," http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1969602

[16] Cho, I-K., Sargent, T. J., Williams, N., 2002. "Escaping Nash Inflation," *Review of Economic Studies* 69, 1-40.

[17] Clauset, A., Shalizi, C. R. and M. E. J. Newman, 2009. "Power-law Distributions in Empirical Data," *SIAM Review* 51(4), 661-703

[18] Collamore, J. F. , 2009. "Random Recurrence Equations and Ruin in a Markov-Dependent Stochastic Economic Environment," *Annals of Applied Probability* 19, 1404–1458.

[19] Cogley, T. and Sargent, T. J., 2008. "The Market Price of Risk and the Equity Premium: A Legacy of the Great Depression?," *Journal of Monetary Economics*, 55, 454–476.

[20] DeJong, D. N. and C. Dave, 2011. Structural Macroeconometrics, 2^{nd} Ed., *Princeton University Press.*

[21] Evans, G., Honkapohja, S., 1999. "Learning Dynamics," *Handbook of Macroeconomics*, Vol.1, eds. J. Taylor and M. Woodford, 1999, North-Holland, pp.449-542.

[22] Evans, G., Honkapohja, S., 2001. *Learning and Expectations in Macroeconomics*. Princeton University Press.

[23] Evans, G., Honkapohja, S., and N. Williams, 2010. "Generalized Stochastic Gradient Learning," *International Economic Review*, 51, 237-262.

[24] Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H. E., 2006. "Institutional Investors and Stock Market Volatility," *Quarterly Journal of Economics*, 121 (2), p. 461-504.

[25] Gabaix, X., 2009. "Power Laws in Economics and Finance," *Annual Review of Economics*, 1, p. 255-93.

[26] Ghosh, A. P, Haya, D., Hirpara, H., Rastegar, R., Roitershtein, A.,Schulteis, A., and Suhe, J, 2010. "Random Linear Recursions with Dependent Coefficients," *Statistics and Probability Letters* 80, 1597 1605.

[27] Holmstrom, B., 1999. "Managerial Incentive Problems: A Dynamic Perspective," *The Review of Economic Studies*, 66, 169-182.

[28] Hollander, F. den, (2000), *Large Deviations*, Fields Institute monographs,American Mathematical Society, Providence, Rhode Island.

[29] Kesten, H., 1973. "Random Difference Equations and Renewal Theory for Products of Random Matrices," *Acta Mathematica*. 131 207–248.

[30] Koulovatianos, C. and V. Wieland, 2011. "Asset Pricing under Rational Learning about Rare Disasters," Manuscript.

[31] Loeve, M. 1977. Probability Theory, 4^{th} Ed., Springer, New York.

[32] Lucas, R. E. Jr., 1978. "Asset Prices in an Exchange Economy," *Econometrica*, Vol. 46, No. 6. (Nov., 1978), pp. 1429-1445.

[33] Nummelin, E., 1984. *General irreducible Markov chains and non-negative operators.* Cambridge Tracts in Mathematics 83, Cambridge University Press.

[34] Petritis, D., *Markov Chains on Measurable Spaces.* Université de Rennes, UFR Mathématiques. perso.univ-rennes1.fr/dimitri.petritis/.../markov/markov.pdf.

[35] Roitershtein, A., 2007. "One-Dimensional Linear Recursions with Markov-Dependent Coefficients," *The Annals of Applied Probability*, 17(2), 572-608.

[36] Saporta, B., 2005. "Tail of the Stationary solution of the Stochastic equation $Y_{n+1} = a_n Y_n + \gamma_n$ with Markovian Coefficients," *Stochastic Processes and their Applications*, 115(12), 1954-1978.

[37] Saporta, B. and Yao, J-F, 2005, "Tail of a Linear Diffusion with Markov Switching," *The Annals of Applied Probability*, 992–1018.

[38] Sargent, T. J., 1999. *The Conquest of American Inflation*. Princeton University Press.

[39] Sargent, T. J. and Williams, N., 2005. "Impacts of Priors on Convergence and Escape from Nash Inflation," *Review of Economic Dynamics,* 8(2), 360-391.

[40] Shiller, R. J., 1999, *Market Volatility*, 6th printing, MIT Press, Cambridge.

[41] Shiller, R. J., 2005. *Irrational Exuberance*, 2^{nd} edition, Broadway Books.

[42] Timmermann, A. 1993, "How Learning in Financial markets Generates Excess Volatility and Predictability in Stock Prices," *Quarterly Journal Economics,*108,1135–1145.

[43] Timmermann, A. 1996, "Excess Volatility and Predictability of Stock Prices in Autoregressive Dividend Models with Learning," *Review of Economic Studies*, 63, 523–557.

[44] Weitzman, M. L. 2007. "Subjective Expectations and Asset-Return Puzzles," *American Economic Review*, 97, 1102–1130.

7. Data Appendix

1. Annual S&P 500 Dataset from Professor Shiller's website (see Shiller (2005))

   (a) The following time series are extracted/constructed for 1871 through 2009 (note that $t = 1, \ldots T$ where $T = 2009.12$):

      i. Extract S & P Comp ($\widetilde{P}(t)$).

      ii. Extract Dividend ($\widetilde{D}(t)$).

      iii. Extract Consumer Price Index ($CPI(t)$).

      iv. Construct Real Price ($P(t)$) as $P(t) = [\widetilde{P}(t) \times\ CPI(T)]/CPI(t)$.

      v. Construct Real Dividend ($D(t)$) as $D(t) = [\widetilde{D}(t) \times\ CPI(T)]/CPI(t)$.

   (b) Construct the Price to Dividends Ratio (ratio) as $P(t)/D(t)$.

2. Quarterly CRSP Dataset

   (a) Download the quarterly data from http://scholar.harvard.edu/campbell/data accessed from Professor Campbell's website, where the particular data being used is associated with "Replication Data for: Consumption Based Asset Pricing". The relevant file is titled USAQE.ASC, note that this is effectively a CRSP dataset with the relevant variables being VWRETD and VWRETX. The text below is an extract from the explanations for this dataset on the above website.

   (b) The following quarterly time series are extracted/constructed for 1926.1 through 1998.4 from the above dataset (note that $t = 1, \ldots T$ where $T = 1998.4$):

      i. Extract Col. 2: $\widetilde{P}(t)$. For each month, the price index is calculated as $\widetilde{P}(t) = (VWRETX(t) + 1) \times \widetilde{P}(t-1)$. (Note that time t in this equation is

in months.) The price index for a quarter, as reported in this column, is the price index for the last month of the quarter. The original data, which goes up to 1996.4 was not altered. The new data, which goes up to 1998.4, was created as described here starting from 1997.1.

ii. Extract Col. 3: $\widetilde{D}(t)$. Dividend in local currency, calculated as follows. The dividend yield for each month is calculated as $\widetilde{DY}(t) = [1+VWRETD(t)]/[1+VWRETX(t)]-1$. Note that if the return index is calculated from $VWRETD$ as above, then this formula agrees with the formula for the dividend yield given earlier. As before, the dividend for each month is calculated as $\widetilde{D}(t) = \widetilde{DY}(t) \times \widetilde{P}(t)$. The dividend for a quarter, as reported in this column, is the sum of the dividends for the three months comprising the quarter.

iii. Extract the Consumer Price Index from Shiller's Monthly Data ($CPI(t)$) which is monthly and associate the last month of a quarter as a quarterly $CPI(t)$.

iv. Construct Real Price ($P(t)$) as $P(t) = [\widetilde{P}(t) \times\ CPI(T)]/CPI(t)$. Take the last price of a quarter as the annual price.

v. Construct Real Dividend ($D(t)$) as $[\widetilde{D}(t) \times\ CPI(T)]/CPI(t)$ and then take quarterly sums to get $D(t)$ at an annual frequency.

vi. Construct the Price to Dividends Ratio (ratio) as $P(t)/D(t)$.

8. Quarterly Frequency Results

We also considered quarterly versions of the data. Quarterly data for the S&P 500 were constructed from the monthly series reported by Professor Shiller on his website. Quarterly data for the CRSP series were directly available from the data website maintained by Professor Campbell (Campbell (2003)). The plots, analogous to Figures 3 and 4 in the text but for quarterly data are as follows.

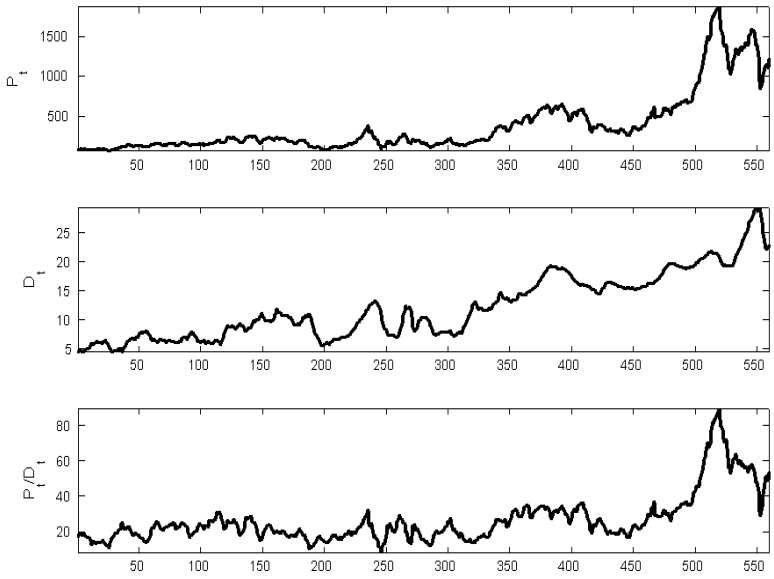

Figure 6. Quarterly S & P 500 (1871-2010).

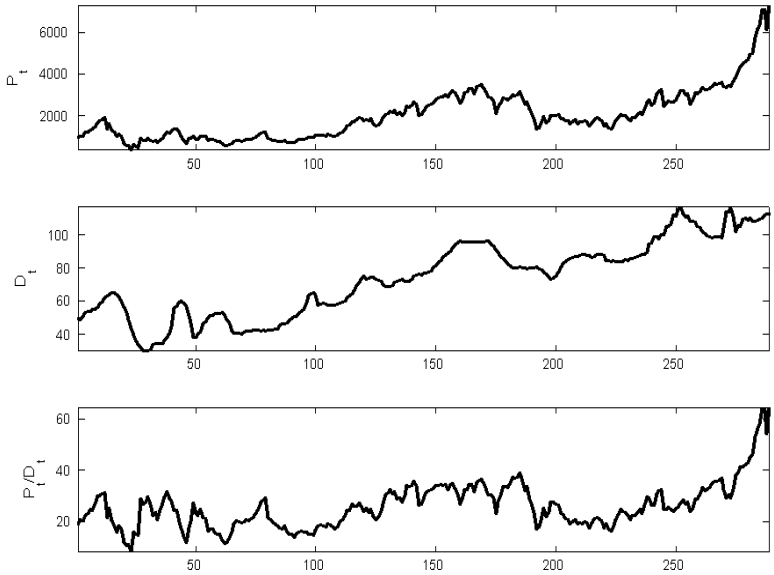

Figure 7. Quarterly CRSP (1926-1998).

Both datasets indicate that at the quarterly frequency the price-dividends ratio is quite volatile, just as in the case with annual data discussed in the text. Next, we employed a baseline parameterization of $(\rho, g, \beta, \gamma) = (0.98, 0.5, 0.99, 2.5)$ consistent with quarterly data (e.g. a $\rho = 0.98$) and conducted the same simulations as reported in the text above, the plots, analogous to Figures 1 and 2 in the text, are as follows.

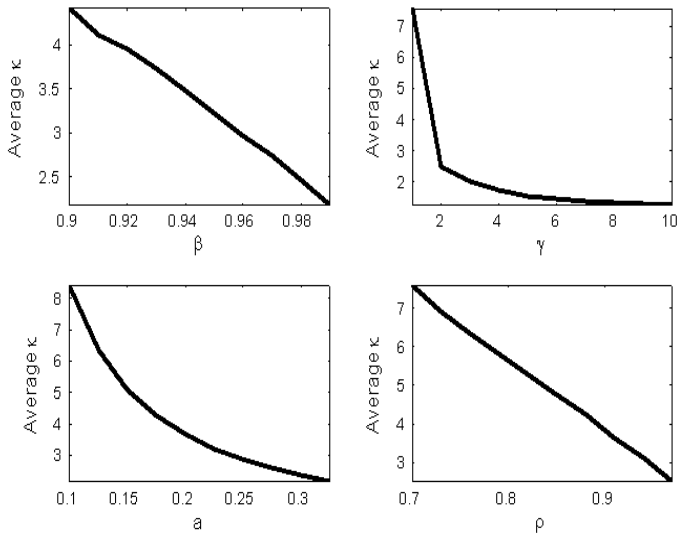

Figure 8. Simulation Results.

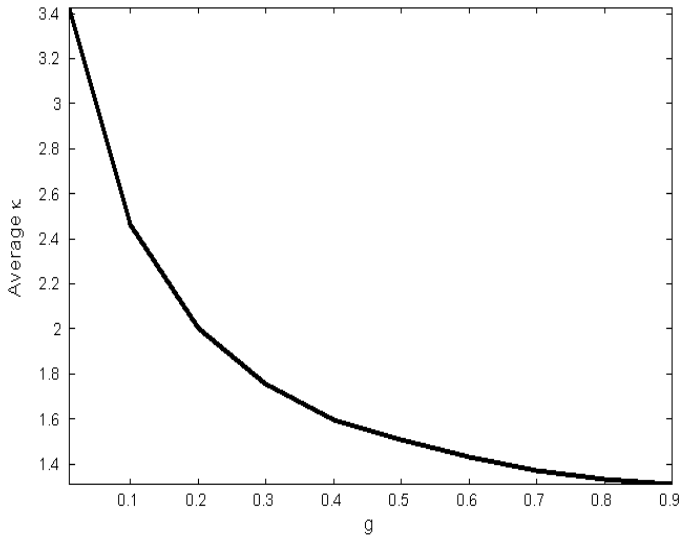

Figure 9. Simulation Results (cont'd.).

Next, our minimum distance estimates analogous to those Tables 1-3 in the text, except employing quarterly data, are as follows.

Table 4. Data Characteristics (Quarterly Data)

	S & P 500	CRSP
	1871QI-2010QIV	1926QIV-1998QIV
$\widehat{\kappa}$	3.5800	6.9894
$s.e.(\widehat{\kappa})$	0.2695	1.3133
$\widehat{\rho}$	0.9826	0.9749
$s.e.(\widehat{\rho})$	0.0078	0.0126
Mean (P_t/D_t)	26.5882	26.0243
Std. Dev. (P_t/D_t)	13.7369	8.7640
Corr (P_t/D_t)	0.9882	0.9456
$r = \frac{\overline{D}_t}{\overline{P}_t}$	0.0322	0.0363
$\beta = (1 + r/4)^{-1}$	0.9920	0.9910
σ_d	0.1836	0.1627

Table 5. Parameter Estimates (Quarterly Data)

	S & P 500		CRSP	
	1871QI-2010QIV		1926QIV-1998QIV	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
g	0.1585	0.0249	0.3746	0.1737
γ	9.1957	0.1194	1.3088	0.1738
β	0.9975	0.0008	0.9974	0.0001
ρ	0.9895	0.0015	0.9592	0.0002
Associated κ	3.5800		6.7959	

Table 6. Drifting Beliefs Model Parameter Estimates

	S & P 500		CRSP	
	1871QI-2010QIV		1926QIV-1998QIV	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
σ_Λ	0.3603	0.1550	1.0114	0.3432
σ_ξ	0.3655	0.0431	0.3032	0.0819
$\log L$	-142.2726		-38.9385	
σ_d	0.1836		0.1627	
Associated g	0.1810		0.5427	

The main results from employing data at a quarterly frequency are twofold. First, while the gain estimates reported in Table 5 fall, it still remains the case that the data exhibit fat tails. This is expected since the data are now at a quarterly frequency. A higher gain or shorter memory with annual data does indeed correspond to a lower gain and longer memory with quarterly data since the definition of the learning horizon (the inverse of the

gain) changes from years to quarters. Second, employing data at a quarterly frequency leads to the structural estimates (as reported in Table 5) being estimated with greater precision. In fact with more data the surface over which a minimum is sought in the minimum distance estimation procedure is sharply defined, leading to the increased precision of the estimates. In contrast to Table 5 however, for Table 6 the CRSP data employed in the drifting beliefs specification lead to a larger than expected gain. This is driven by a much higher estimate for σ_Λ relative to the annual data.

Finally, we also implemented a block bootstrap in estimating the drifting beliefs model. That is, we used a block bootstrap to generate 1000 samples from the data, each of which was used to estimate the parameters of a drifting beliefs model with an attendant computation of g from equation (44), for example, as reported in the last row of Table 6. This yielded 1000 estimates of g for which we then computed the mean and standard deviation. Such estimates yielded a mean (standard deviation) of g of 0.5080 (0.1777) using quarterly CRSP data and a mean (standard deviation) of g of 0.6114 (0.1362) when using S&P 500 data. With annual data the estimates of g were 0.2095 (0.1788) using CRSP data and 0.2196 (0.1788) using S&P 500 data. Thus, with quarterly data using the minimum distance method or, using bootstraps with either quarterly or annual data, the estimates of g were sharper and the standard errors fell.