Learning, Large Deviations and Rare Events

Jess Benhabib
NYU

Chetan Dave
NYU

Introduction

- The learning literature (e. g. Marcet and Sargent (1989), Woodford (1990) and Evans and Honkapohja (2001)) replaced expectations with regressions in dynamic stochastic models.
- A main focus was to explore how 'learning' $\longrightarrow$ convergence to rational expectations equilibria (REE) in dynamic stochastic economic models.

- The idea that learning processes can generate volatility has been explored in the asset pricing literature (e. g. Timmerman (1993, 1996, 2007),Weitzman (2007), Adams, Marcet and Nicolini (2006)).
- Sargent (1999) and Cho, Sargent and Williams (2002), Sargent and Williams (2005) introduced a new idea: Under recursive least squares constant gain (RLSCG) learning algorithms, uncertainty about estimated parameters can persist and fuel 'escape' dynamics in which a sequence of rare and unusual shocks propel agents away from the REE of a model (see also Williams (2009)): They suggested that this can be studied through "large deviations theory".
- In the context of monetary policy with learning, Cho, Sargent and Williams (2002) showed via simulations that escape dynamics can fuel large deviations in inflation.

Constant gains and optimal learning

- Sargent and Williams (2005) incorporated the expectation on the part of agents that the estimated parameters follow a random walk, so that uncertainty about parameters persists over time. (We will see that this is a natural assumption under adaptive learning)) They then show that the generalized constant gain stochastic gradient (SGCG) algorithm is the optimal Bayesian estimator in that case. (See also Holmstrom (1999)). 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.
- So "escapes" or "large deviations" in economic variables can take place when sequences of large shocks throw off the learning process from the rational expectations equilibrium.

Constant gains and optimal learning

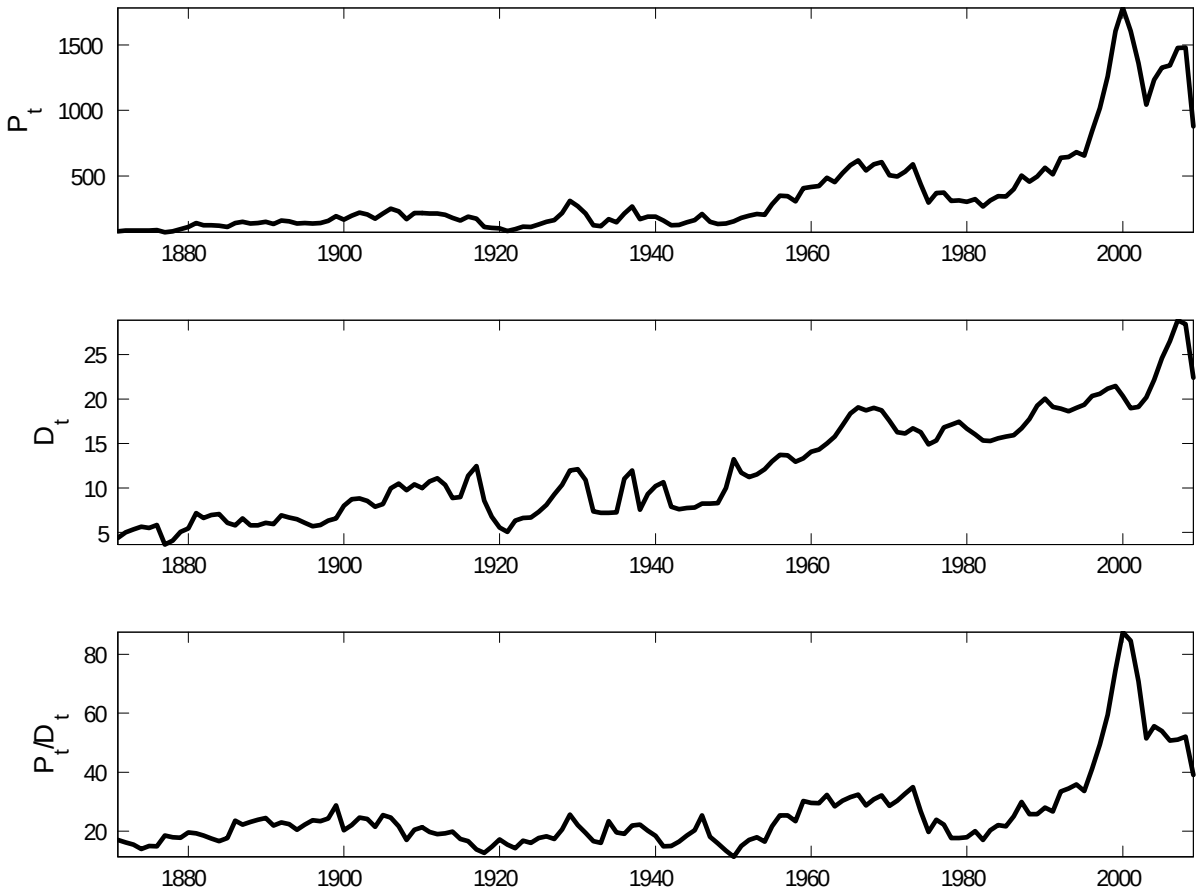
- But how big should the constant gain be? It will depend on the underlying variances of the exogenous driving process (dividends), the perceived variance of the randow walk, and the variance of the agent's "perceived law of motion". Will the constant gain be self-confirming?

Introduction

- We characterize the limiting probabilities of escape dynamics and large deviations from a REE under Constant Gain Stochastic Gradient learning.
- **Context**: a univariate linear expectational difference equation
  - Encompasses asset pricing, overlapping generations models and others (see Evans and Honkapohja (1999)).
- **Finding**: Recursion for estimated coefficient can exhibit occasional large deviations (rare events) inducing a limiting 'fat tailed' power law distribution.
  - **Under Asset Pricing Interpretation**: Ratios of asset prices to dividends can exhibit volatility and deviate from their REE values.

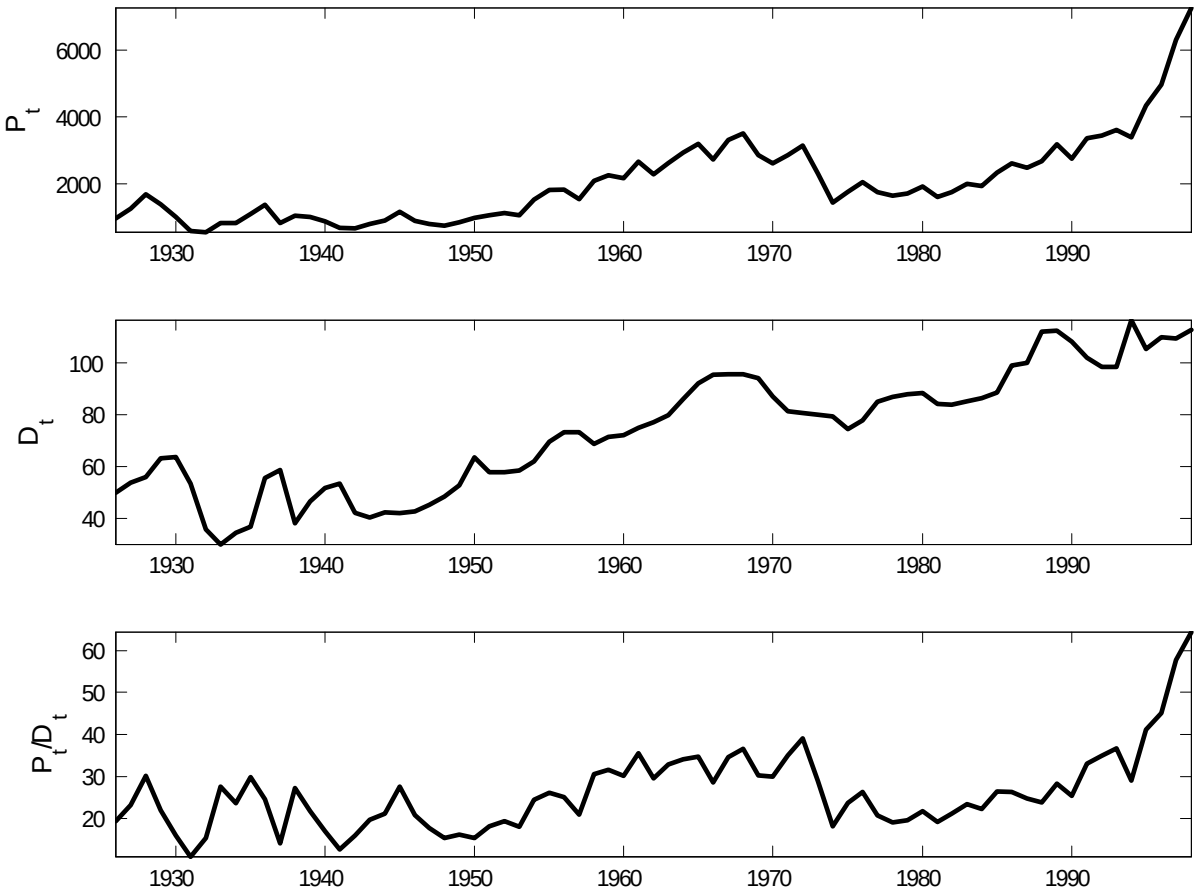
Asset Prices

Figure 1. Monthly S & P 500, 1871.1 - 2010.12, (Source: Shiller).



Asset Prices

Figure 2. Quarterly CRSP, 1926.1 - 1998.4, (Source: Campbell (2003)).



The Model

- Consider

$$p_t = \delta E_t(p_{t+1}) + \chi d_t, \quad \delta \in (0,1). \tag{1}$$

with d_t an exogenous process

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

where $\varepsilon_t \sim i.i.d.(0, \sigma^2)$ with compact support $[-a, a]$, $a > 0$.

The Model

- **Asset Pricing Interpretation**: Consider a single asset in a Lucas (1978)-type economy, CRRA preferences parameterized by γ, then

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

yields

$$p_t = \delta E_t(p_{t+1}) + \chi d_t, \quad \delta \in (0,1), \quad \chi \equiv (1-\delta-\gamma)\rho + \gamma$$

for $p_t = \log(P_t) - \log(\overline{P})$ and $d_t = \log(D_t) - \log(\overline{D})$.

The Model

- The REE of the model in (1)-(2) is

$$p_t = \phi d_{t-1} + \eta_t, \quad \phi = \frac{\chi\rho}{1-\delta\rho}, \quad \forall\ \delta\rho \neq 1. \tag{3}$$

- Under learning, agents assumed to be learning ϕ. First they form a perceived law of motion (PLM)

$$p_t = \phi_{t-1} d_{t-1} + \xi_t \tag{4}$$

where ξ_t is a regression error and estimates of ϕ_{t-1} will have a recursive form.

- Given the PLM, agents *assumed* to form expectations as

$$E_t(p_{t+1}) = \phi_{t-1} d_t. \tag{5}$$

The Model

- Next, insert $E_t(p_{t+1}) = \phi_{t-1} d_t$ into (1) to obtain actual law of motion (ALM)

$$p_t = (\delta\phi_{t-1} + \chi)\rho d_{t-1} + (\delta\phi_{t-1} + \chi)\varepsilon_t, \tag{6}$$

- **Aside**: Note how the fixed point of the map from PLM to the ALM ($\phi = (\delta\phi + \chi)\rho = T(\phi)$) will deliver the REE (this is the T-map of Evans and Honkapohja (2001)).
- *Assumed* timing within period t:
  - **1.** agent forms $E_t(p_{t+1})$ using ϕ_{t-1} and d_t
  - **2.** p_t realized from ALM.
  - **3.** agent estimates ϕ_t using all data including p_t

The Model

- Recall the PLM

$$p_t = \phi_{t-1} d_{t-1} + \xi_t, \qquad \xi_t \sim iid(0, \sigma^2_\xi),\ \sigma^2_\xi < +\infty \tag{7}$$

- Given Sargent and Williams (2005) and Evans et al. (2010), we focus on a constant gain stochastic gradient learning algorithm to update ϕ_t as

$$\phi_t = \phi_{t-1} + g d_{t-1}(p_t - d_{t-1}\phi_{t-1}), \quad g \in (0, 1) \tag{8}$$

where g is the gain parameter. Note that under constant gains, from the perspective of agents ϕ_t is a random walk and from the PLM $(p_t - d_{t-1}\phi_{t-1})$ is a zero mean forecast error.

The Model

- What is g if agents try to compute it? In the limit, as data accumulates, the optimal g from the Kalman filter, if agents believe ϕ_{t-1} is a random walk
  $\phi_t = \phi_{t-1} + \Lambda_t,\ \Lambda_t \sim iid(0, \sigma_\Lambda^2),\ \sigma_\Lambda^2 < +\infty$,then

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

  where σ_d is the variance of the stationary distribution of dividends $\{d\}$. We will come back to this point.

Characterizing Large Deviations

- Insert the ALM in place of p_t in (8) to obtain

$$\phi_{t+1} = \lambda_{t+1}\phi_t + \psi_{t+1} \tag{9}$$

with

$$\lambda_{t+1} = 1 - (1 - \rho\delta)gd_t^2 + \delta g d_t \varepsilon_{t+1} = 1 - gd_t^2 + g\delta d_{t+1} d_t$$

and

$$\psi_{t+1} = \chi\rho g d_t^2 + \chi g d_t \varepsilon_{t+1} = \chi g d_{t+1} d_t. \tag{10}$$

- Now $\{\lambda_t\}$ generates multiplicative noise for $\{\phi_t\}$.
- Can be source of large deviations and fat tails for stationary distribution of $\{\phi_t\}$.
- Note:$\{\phi\}_t$ is not a random walk....

Characterizing Large Deviations

- Kesten (1973), Goldie (1991), Roitershtein (2007): Let

$$\phi_{t+1} = \lambda_{t+1}\phi_t + \psi_{t+1} = \phi_1 \prod_{i=2}^{t+1} \lambda_i + \left(\psi_{t+1} + \sum_{j=1}^{t} \psi_j \prod_{i=j+1}^{t+1} \lambda_i \right)$$

and assume $\lambda_t > 0$ has compact support with $E(\lambda_t) < 1$ and $P(\lambda_t > 1) > 0$ at its stationary dist^n. Then $\exists$ a $\kappa > 0$ solving

$$\Lambda(\kappa) = \lim_{t \to \infty} \sup \frac{1}{t} \log E \left(\prod_{i=0}^{t} \lambda_i^{\kappa} \right) = 0$$

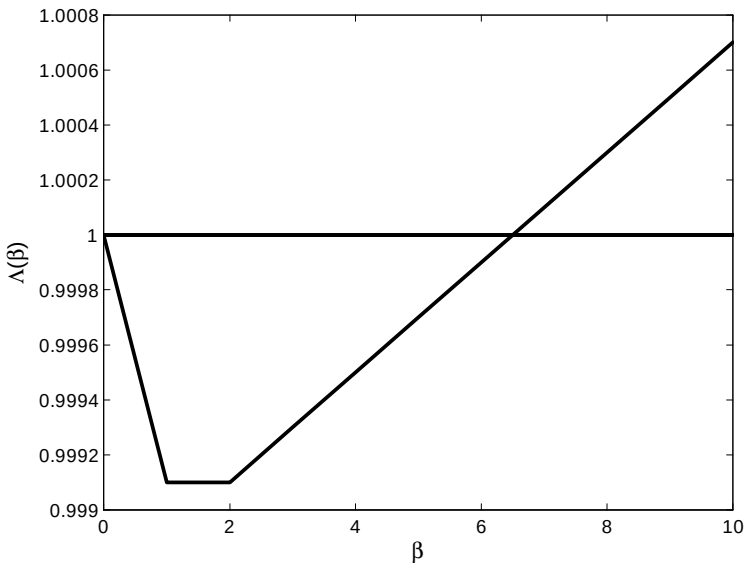
Then, modulo technical assumptions, tails of the stationary distribution of $\{\phi_t\}$ can be characterized by:

$$\lim_{\tau \to \infty} \tau^{\kappa} P(\phi > \tau) = K_1 > 0 \;\&\; \lim_{t \to \infty} \tau^{\kappa} P(\phi > -\tau) = K_{-1} > 0$$

where K_1 and K_{-1} are not both 0.

Characterizing Large Deviations

- As $\kappa \uparrow$, tail of the stationary distribution of $\{\phi_t\}_t$ thins. A (simulated) plot of $\Lambda(\kappa)$ vs. κ is below, showing a particular κ



Characterizing Large Deviations

- To ensure $E\left|\lambda_{\infty}\right|<1$, we restrict support of $\varepsilon_t \in [-a, a]$ by assuming

$$a<\left(\frac{6\left(1-\rho^2\right)}{g\left(1-\beta\rho\right)}\right)^{0.5} \tag{11}$$

- The proof basically follows from showing
  - *i*) $\Lambda(0)=0$, and if $E\left|\lambda_{\infty}\right|<1$, then $\Lambda'(0)<0$,
  - *ii*) *If* $P\left(|\lambda|>1\right)>0$, then $\lim_{\delta\rightarrow\infty}\Lambda(\delta)=\infty$,
  - *iii*) $\Lambda(\kappa)$ is convex.

Characterizing Large Deviations

- Proposition characterizes the tail of the stationary distribution of ϕ as a power tail with exponent κ.
- Distribution of ϕ has moments only up to the highest integer less than κ
    - Therefore is a 'fat tailed' distribution rather than a Normal distribution.
- Results driven by: the stationary distribution of $\{\lambda_t\}_{t\in\mathbb{N}}$ has a mean less than 1 but has support above 1 with positive probability.
- Thus large deviations are strings of realizations of λ_t above one, even though they are rare events, and can produce fat tails.

Characterizing Large Deviations

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

Model Simulations and Comparative Statics

- To 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. We can then explore how our estimate of κ from simulated series varies as we vary parameters.
- We simulate 100 series 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.

Model Simulations and Comparative Statics

- Escapes or large deviations in prices take place when sequences of large shocks to dividends throw off the learning process. Such escapes are more likely if dividend shocks produce larger support of λ_t above 1. We expect lower κ, or fatter tails, as the support of λ_t that lies above 1 gets larger.

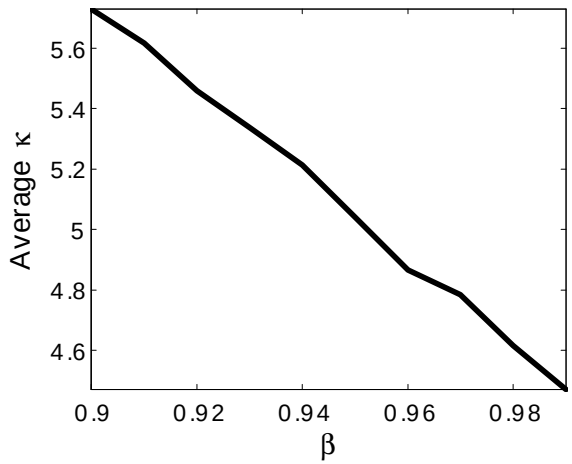
Comparative Statics

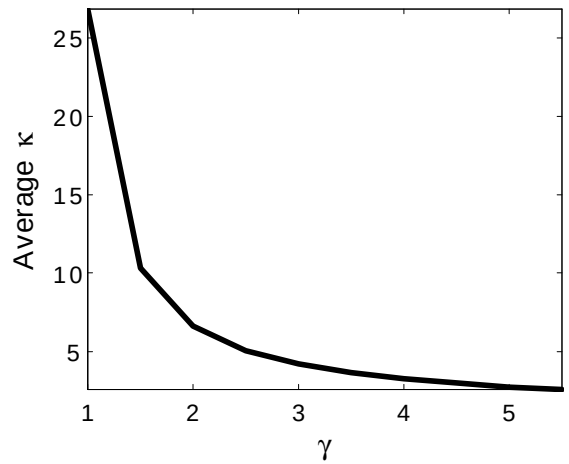
- For simulations, given the parameter estimates in the next section, we use the baseline parameterization, $(\rho, g, \beta, \gamma) = (0.98, 0.5, 0.95, 2.5)$. The restriction given by equation on a implies a maximum value of $a = \hat{a} = 2.6243$, so for the baseline parametrization we set the baseline value of $a = 0.225$.
- We find that the average κ is 4.9172, the average (P_t / D_t) is 20.4989 and the average Std. Dev (P_t / D_t) is 12.6142. This are quite close to the data characteristics in Table 1.

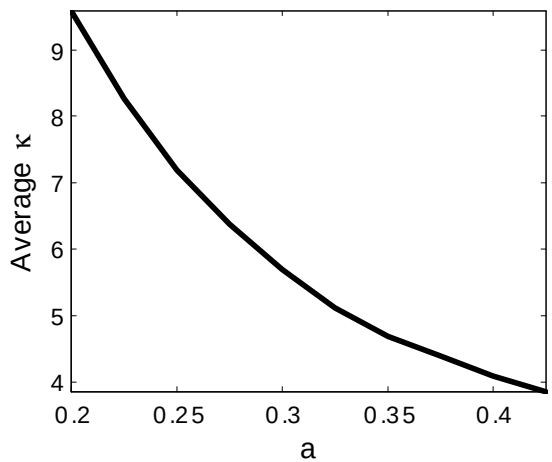
Comparative Statics

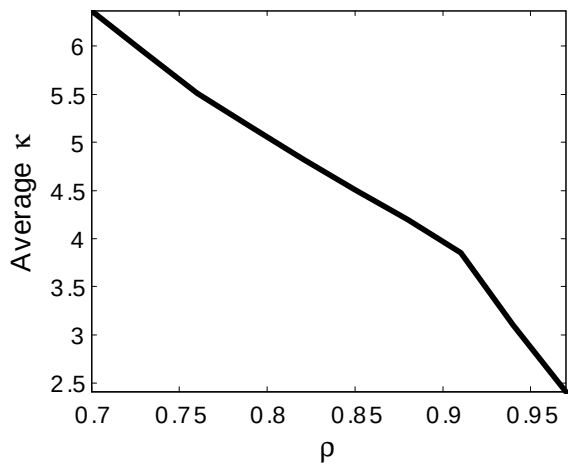
.

χ should be γ)



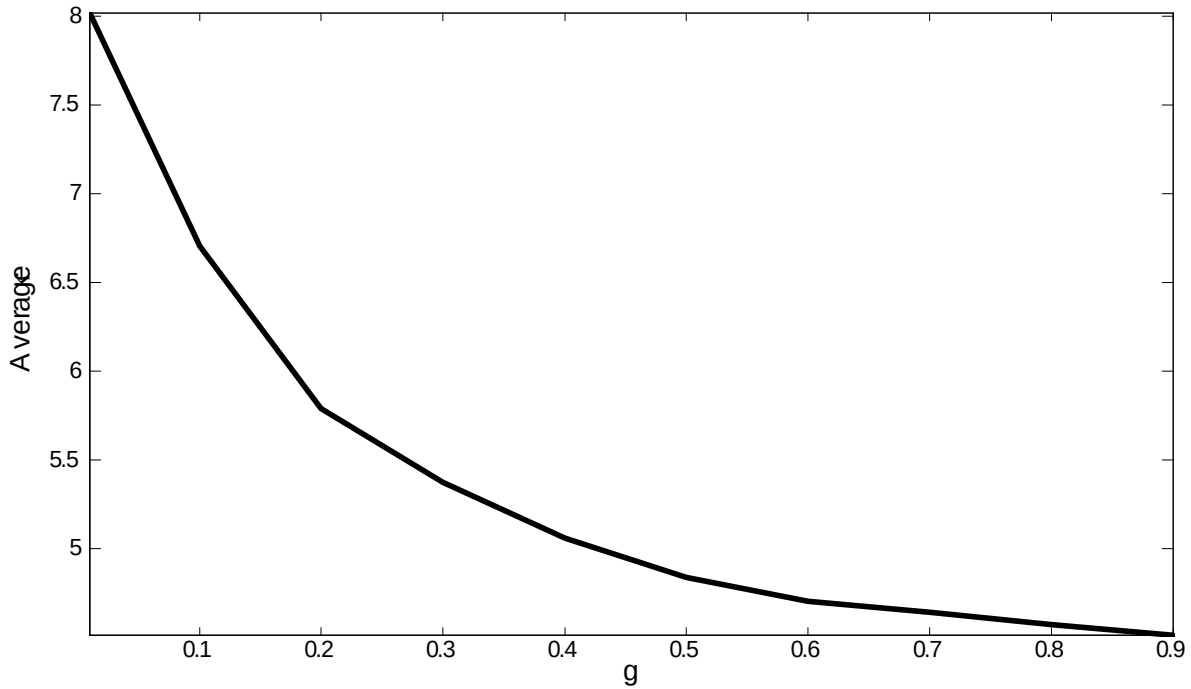






Comparative Statics

Figure 4. Simulation Results (cont'd.).



As the learning gain falls, that is, the horizon for learning increases, average κ rises. However, for empirically plausible values of g, is the average κ small?

Comparative Statics

- This of course is in accord with the Theorem 7.9 in Evans and Honkapohja (2001). As the gain parameter $g \longrightarrow 0$ and $tg \longrightarrow \infty$, $\left\{\phi_t^g - \varkappa\right\} / g^{0.5}$ converges to a Gaussian variable where $\varkappa$ is the globally stable point of the associated ODE $(\dot{\phi} = T(\phi) - \phi)$ describing the mean dynamics.
- More generally, as $g \longrightarrow 0$, the estimated coefficient under learning with gain parameter g, ϕ_t^g, converges in probability (but not uniformly) to $\varkappa$ for $t \longrightarrow \infty$. However, there will always exist arbitrarily large values of t with ϕ_t^g taking values remote from $\varkappa$ (See Benveniste, Métivier and Priouret (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 $g > 0$.

Empirics
Data Characteristics

- We use a maximum likelihood procedure following Clauset et al. (2009) to estimate κ associated with P_t / D_t for both S&P and CRSP dividend series.
- κ is small for both series, suggesting that only the first few moments of P_t / D_t exist irrespective of the data source.
- Table 1 below 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.

Table 1. Data Characteristics

	S & P 500	CRSP
$\widehat{\kappa}$	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

Empirics
Parameter Estimates

- We use two separate approaches to get estimates for the gain parameter g.
- 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

$$\min_{\vartheta} \left[\kappa - \kappa(\vartheta)\right]^2.$$

- Minimization was conducted using a simplex method and standard errors were computed using a standard inverse Hessian method.

Empirics
Parameter Estimates

- The minimization procedure proceeds as follows. For candidate parameterizations of ϑ we employ the S&P or CRSP series dividends d_t to calculate ϕ_t. The ALM 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.
- Table 2 below reports the estimates $(\widehat{\vartheta})$ and associated standard errors $(s.e.(\widehat{\vartheta}))$ for each of the the S&P or CRSP dividend series, as well as the κ associated with the estimated parameters.

Empirics

Parameter Estimates-First approach

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	

Empirics
Parameter Estimates

- Point estimates of g plausible, although standard errors are quite large. Carceles-Poveda and Giannitsaraou (CPG) (2008) discuss plausible values of g, where under constant gain the decay in weights on past observations dating i periods back is $(1-g)^{i-1}$.
- For quarterly observations, $g = 0.46$ corresponds to 15 years of learning, with periods beyond 15 years having practically zero weight (but for CPG practically zero is very very small). For learning going back 20 years, $g = 0.37$.
- CPG (2008) report that the standard deviations of price-dividend ratios for the Lucas asset pricing model under rational expectations or learning are smaller than the standard deviations in the data by factors of about 20 to 50.

Empirics
Parameter Estimates

- Note that our estimates match the parameter values used by Carceles-Poveda and Giannitsaraou (2008) in their simulations except for γ, the CRRA parameter, which they set equal to 1 while we have it at $\gamma = 2.5$. Note also from comparative statics above that κ drops dramatically with γ .

	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	

Empirics
Parameter Estimates-Second Approach

- The PLM is

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

- ϕ follows a random walk:

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

- The Bayesian agent would use these to estimate σ_Λ, σ_d and σ_ξ to set an optimal estimate of the gain in the limit as

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

Empirics

- To compute g an estimate of σ_d is of course readily obtained from the actual 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 using actual price data, and employing the methods laid out in Kim and Nelson (1999) we can obtain estimates of σ_Λ and σ_ξ.

Table 3. 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	

- g is much larger than what is usually assumed in the literature. In Figure 4, a value of $g = 0.4866$ yields a tail κ of about 4.9 while $g = 0.5455$ 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 parameters and gains of 0.4866 and 0.5455. These resulted in average price-dividend ratios of 20.6324 and 20.6965 with standard deviation values of 10.0051 and 10.5870.

Empirics

- 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 using the methods in Kim and Nelson (1999). The average g is 0.3826, which is quite close to and confirms the benchmark value of $g = 0.4$ that is used in generating the simulated data.
- Question: Are there clever econometric methods for the agent to test whether $\{\phi\}_t$ is not a random walk? They can test the mean of $\{\phi_t - \phi_{t-1}\}$ to see if they can reject if is zero. They can check if its variance matches the forecast error in the PLM?

Conclusion

- Asks whether agents 'learn' REE.
- Constant gain learning: heavier emphasis on recent observations, can lead to escape dynamics.
- Escape dynamics can propel estimated coefficients away from the REE values.
- We show (under asset pricing interpretation), 'bubbles,' or asset prices that exhibit large deviations from their REE ratios to dividends can occur with a frequency associated with a fat tailed power law.
- Techniques can be generalized to higher dimensions and to finite state Markov chains, can be applied to other more general economic models.