APPENDIX – FOR ONLINE PUBLICATION

A Appendix Figures

Figure A.1: Estimation Error: Error-Revision Coefficient and Implied Persistence Coefficient

This figure shows simulation results on the error-revision coefficient and the implied persistence coefficient. We start by simulating 10 datasets of 45 participants each, where each participant makes 40 forecasts of an AR(1) process. Each of the 10 dataset has one level of the AR(1) persistence ρ, which goes from 0 to 1. In each dataset, participants make forecasts using the diagnostic expectations model: $F_t x_{t+h} = \rho^h x_t + 0.4\rho^h \epsilon_t$, where x_t is the process realization and ϵ_t is the innovation. In panel A, for each level of ρ, we estimate the error-revision coefficient b from the following regression: $x_{t+1} - F_t x_{t+1} = c + b(F_t x_{t+1} - F_{t-1} x_{t+1}) + u_{t+1}$. The dark solid line shows the theoretical prediction (Bordalo et al., 2020c). The light solid line shows the average coefficient from 200 simulations. The dashed lines show the 90% confidence bands from the simulations. In Panel B, we implement the same procedure and report the implied persistence coefficient $\hat{\rho}$ estimated from the regression: $F_t x_{t+1} = cst + \hat{\rho} x_t + v_{t+1}$. The dark solid line shows the theoretical prediction based on diagnostic expectations. The light solid line shows the average coefficient from 200 simulations. The dashed lines show the 90% confidence bands from the simulations. The standard errors are very tight so the three lines lie on top of one another.

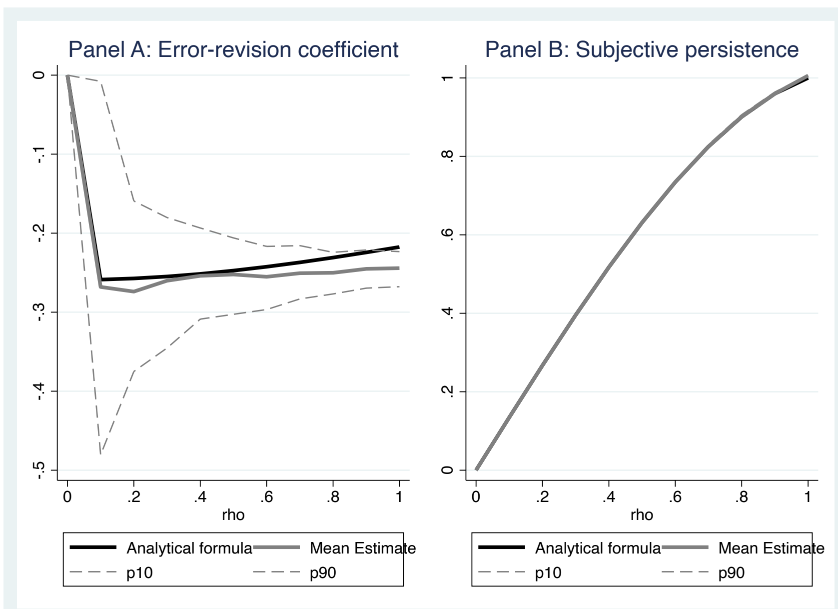


Figure A.2: Prediction Screen

This figure shows a screenshot of the prediction task. The green dots indicate past realizations of the statistical process. In each round t, participants are asked to make predictions about two future realizations $F_t x_{t+1}$ and $F_t x_{t+2}$. They can drag the mouse to indicate $F_t x_{t+1}$ in the purple bar and indicate $F_t x_{t+2}$ in the red bar. Their predictions are shown as yellow dots. The grey dot is the prediction of x_{t+1} from the previous round ($F_{t-1} x_{t+1}$); participants can see it but cannot change it. After they have made their predictions, participants click “Make Predictions" and move on to the next round. The total score is displayed on the top left corner, and the score associated with each of the past prediction (if the actual is realized) is displayed at the bottom.

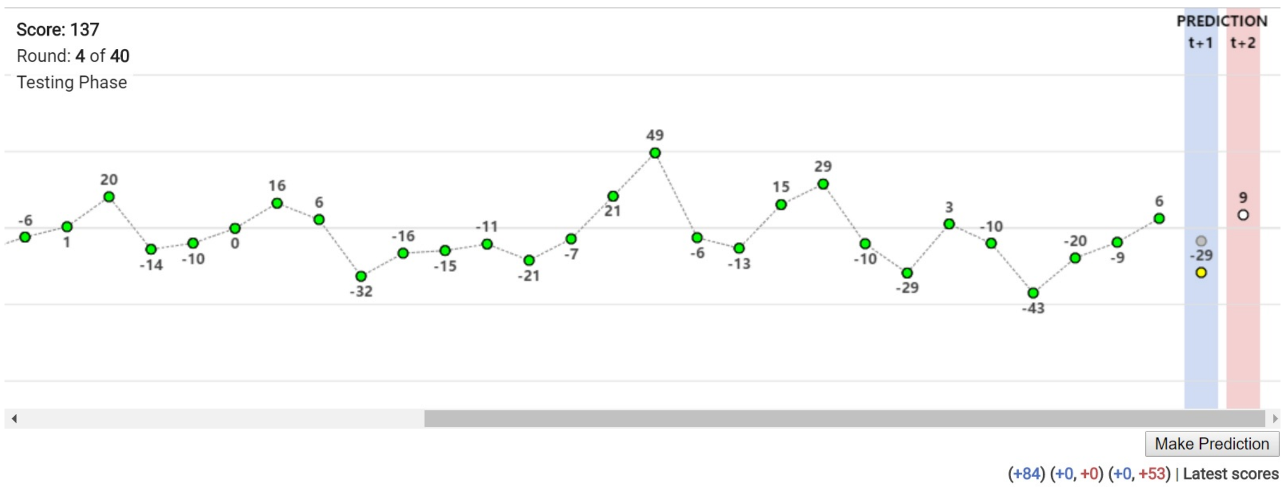


Figure A.3: Implied Persistence and Actual Persistence

We compute the implied persistence ρ_1^s from $F_{it} x_{t+1} = c + \rho_1^s x_t + u_{it}$ for each level of AR(1) persistence ρ. The y-axis plots the implied persistence relative to the actual persistence $\zeta = \rho_1^s / \rho$, i.e., the measure of overreaction, and the x-axis plots the AR(1) persistence ρ. The line at one is the FIRE benchmark.

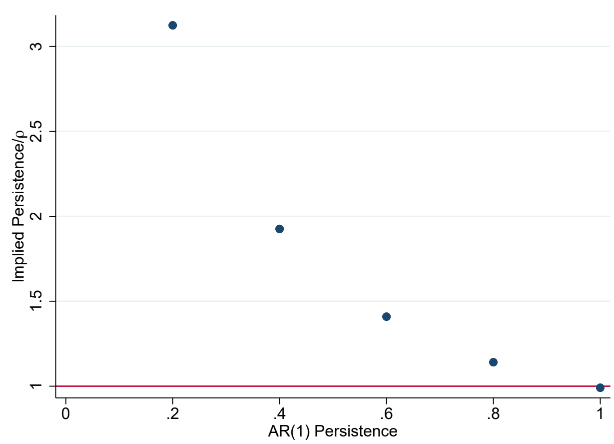
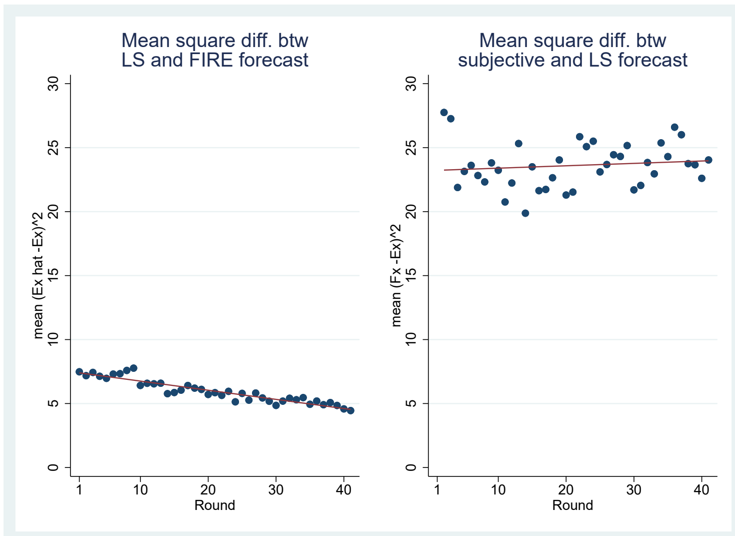


Figure A.4: Distance between Subjective Forecasts and Rational Expectations

The top left panel shows the root mean squared difference between in-sample least square expectations and full information rational expectations (FIRE). The top right panel shows the root mean squared difference between participants' actual subjective forecasts and the least square forecasts. The data use all conditions in Experiment 1. The bottom panel shows the implied persistence of least square forecasts for each level of ρ, which is the regression coefficient of the least square forecast on x_t.

Panel A. Least Square Forecasts vs. FIRE and Subjective Forecasts



Panel B. Implied Persistence of Least Square Forecasts

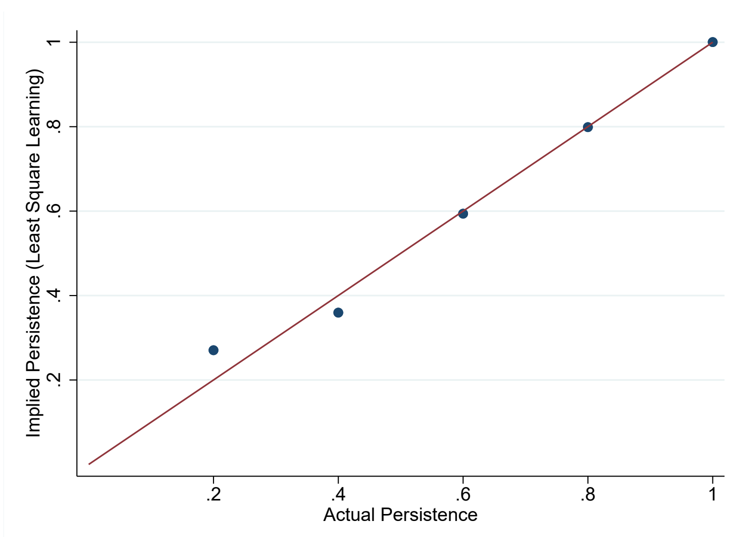
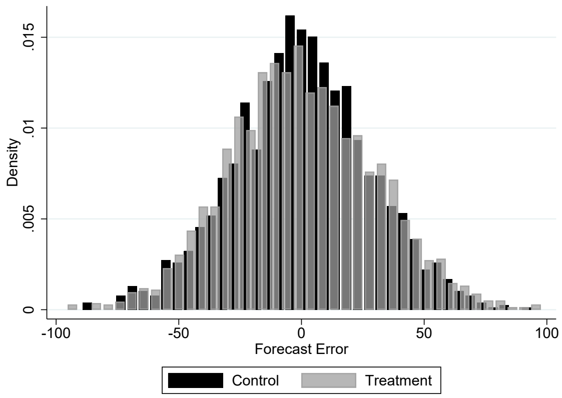


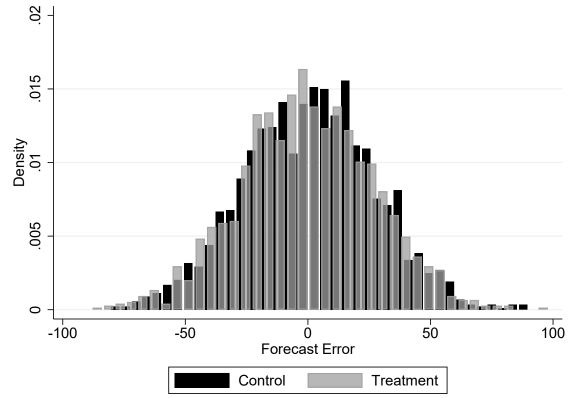
Figure A.5: Knowledge of Linear DGP and the Distribution of Forecasts

We use the data from Experiment 3 (MIT EECS), with 204 MIT undergraduates randomly assigned to AR(1) processes with $\rho = .2$ or $\rho = .6$. 94 randomly selected participants were told that the process is a stable random process (control group), while 110 were told that the process is an AR(1) with fixed μ and ρ (treatment group). Panel A shows the distributions of the forecast error $x_{t+1} - F_t x_{t+1}$ for both treated and control groups. Panel B shows binscatter plots of the forecast error as a function of the latest realization x_t.

Panel A. Distribution of Forecast Error ($x_{t+1} - F_t x_{t+1}$)

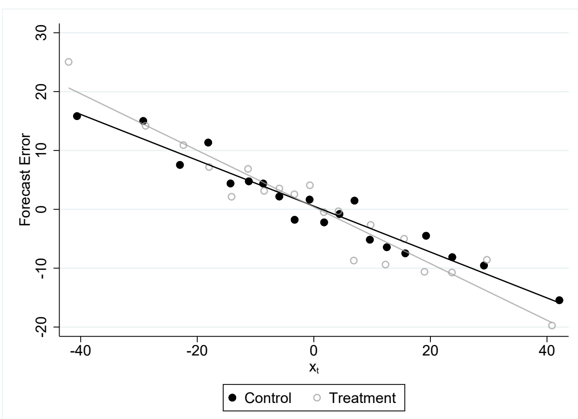


$\rho = 0.2$

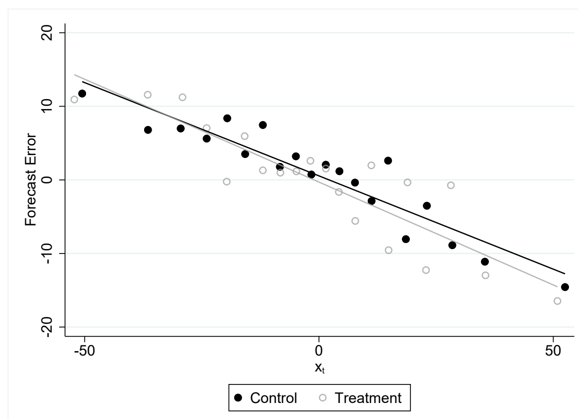


$\rho = 0.6$

Panel B. Forecast Error Conditional on x_t



$\rho = 0.2$

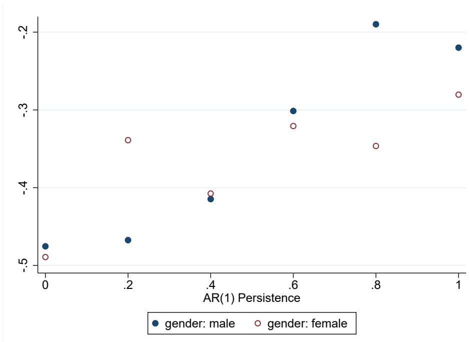


$\rho = 0.6$

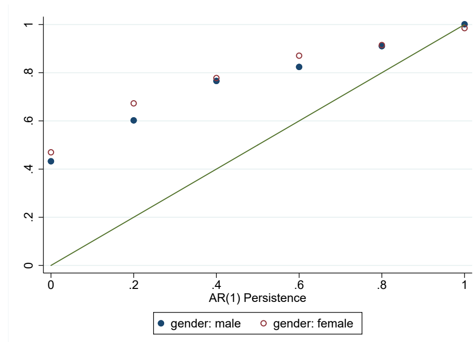
Figure A.6: Overreaction and Persistence of Process: Results by Demographics

This figure plots the error-revision coefficient and the implied persistence for each level of AR(1) persistence, estimated in different demographic groups. In Panel A, the solid dots represent results for male participants and the hollow dots represent results for female participants. In Panel B, the solid dots represent results for participants younger than 35 and the hollow dots represent results for participants older than 35. In Panel C, the solid dots represent results for participants with high school degrees, and the hollow dots represent results for participants with college and above degrees.

Panel A. By Gender: Male vs. Female

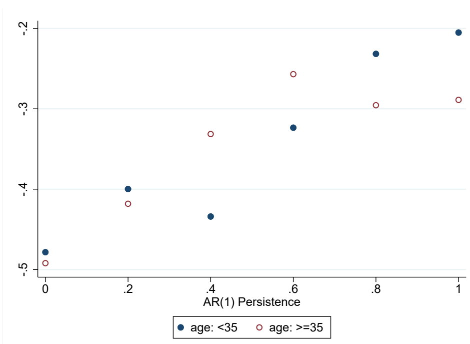


Error-revision coefficient by Gender

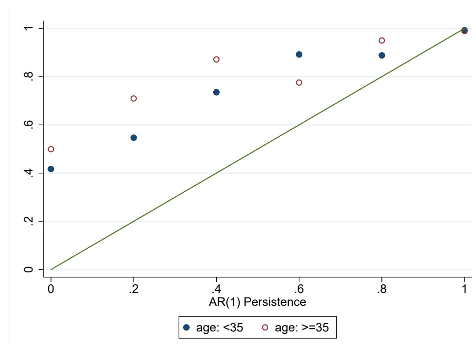


Implied Persistence by Gender

Panel B. By Age: Below 35 vs. Above 35

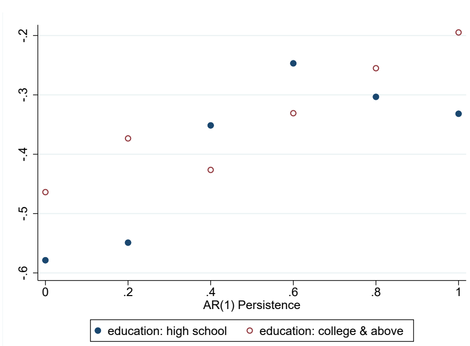


Error-revision coefficient by Age

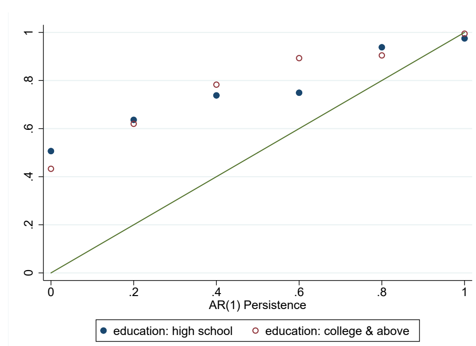


Implied Persistence by Age

Panel C. By Education: High School vs. College and Above



Error-revision coefficient



Implied Persistence

Figure A.7: Error-Revision Coefficient: Data vs Models

For each level of ρ, we regress the model-based forecast error $x_{t+1} - \widehat{F_t^m x_{t+1}}$ on the model-based forecast revision $\widehat{F_t^m x_{t+1}} - \widehat{F_{t-1}^m x_{t+1}}$. The dots report the regression coefficient obtained for each model m and each level of ρ. The solid line reports the error-revision coefficient in the experimental data, as in Figure II, Panel A.

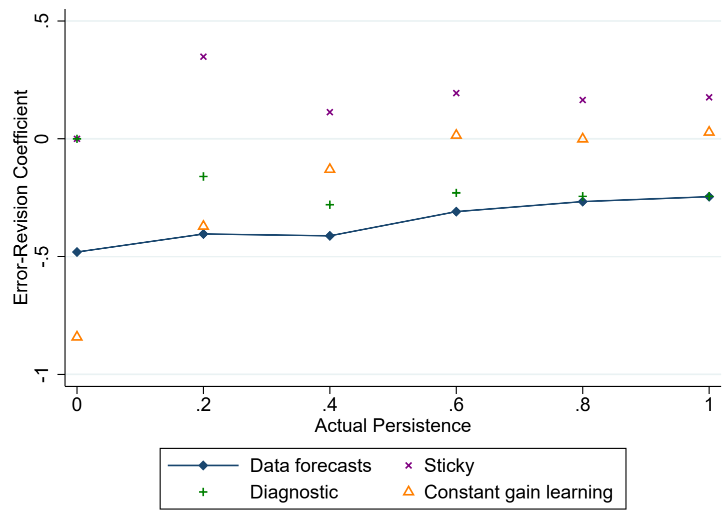


Figure A.8: Implied Persistence and Forecast Horizon

We report the forecast implied persistence for different horizons. For each horizon h and a given ρ, the x-axis is based on ρ^h, and the y-axis shows ρ_h^s which is the regression coefficient of $F_t x_{t+h}$ on x_t. Full dots correspond to $h = 1$ (from Experiment 1 where the one-period persistence $\rho \in \{0, .2, .4, .6, .8, 1\}$), which are identical to Figure II, Panel B. Empty circles correspond to $h = 2$ (also from Experiment 1). Crosses correspond to $h = 5$ and come from Experiment 3, where the one-period persistence $\rho \in \{.2, .4, .6, .8\}$. The solid orange line is the 45-degree line.

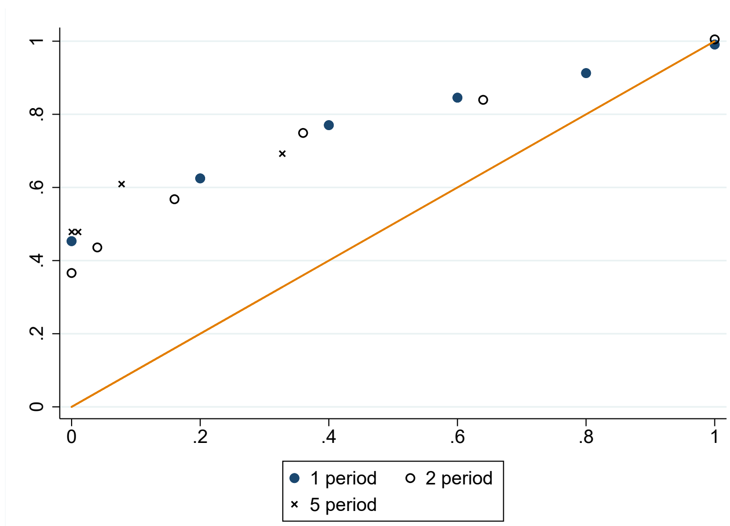
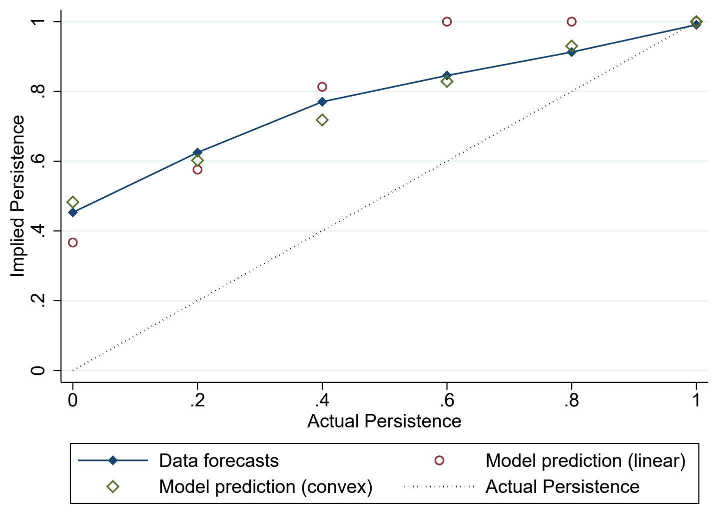


Figure A.9: Model Functional Form: Robustness Checks

This figure shows the model fit under alternative model specifications of the cost function, for $h = 1$ in Panel A and $h = 5$ in Panel B. The red dots represent the implied persistence from our model when $\gamma = 1$, and the green diamonds represent result from our model when we do a full grid search for γ. The blue line represents the value observed in the forecast data.

Panel A. $h = 1$



Panel B. $h = 5$

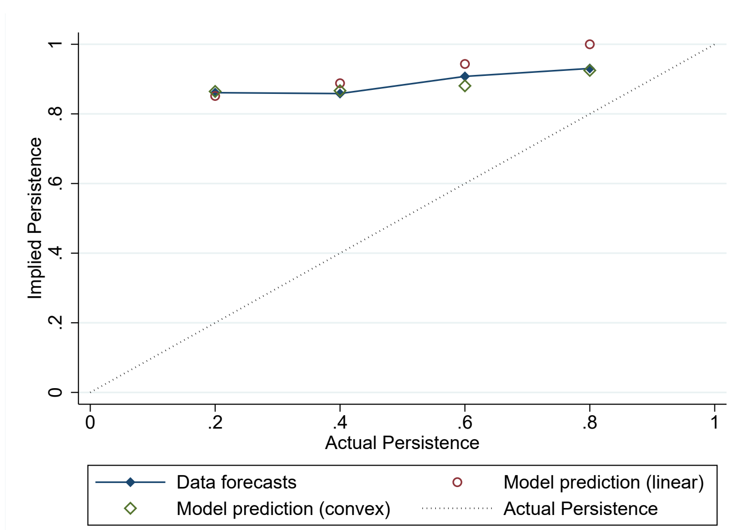
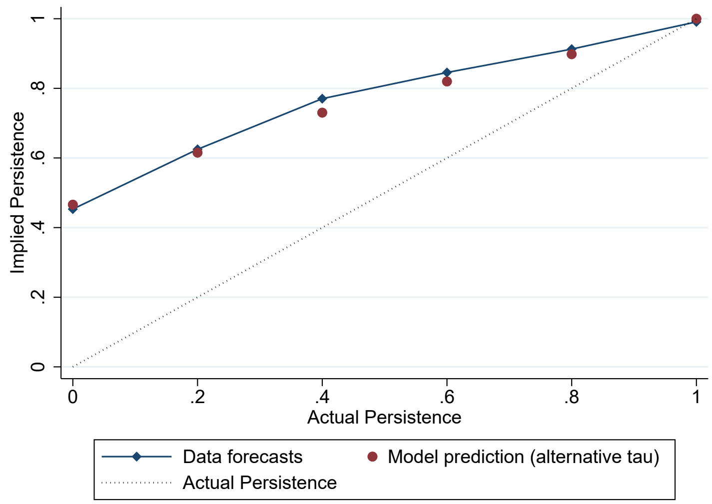


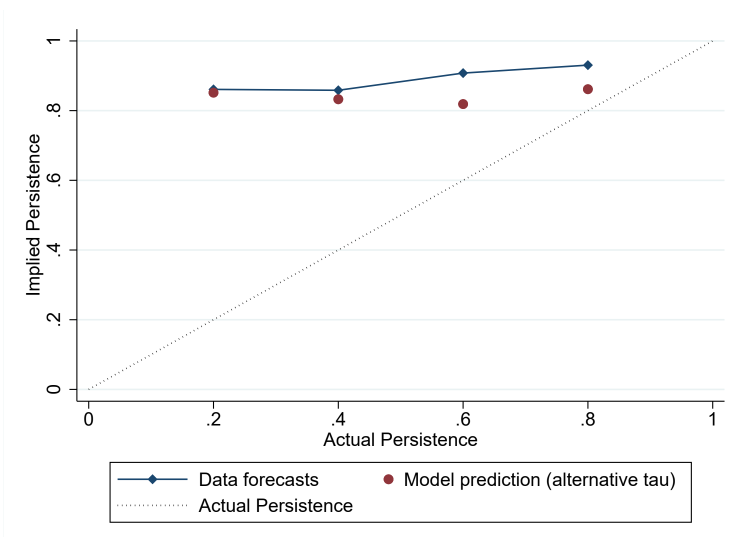
Figure A.10: Model Functional Form: Robustness Checks

This figure shows the model fit under the alternative formulation of $\underline{\tau}$, as discussed in Section 6.2, for $h = 1$ in Panel A and $h = 5$ in Panel B. The red dots represent the implied persistence from our model, and the blue line represents the value observed in the forecast data.

Panel A. $h = 1$



Panel B. $h = 5$



B Appendix Tables

Table A.1: Experimental Literature on Expectations Formation

This table summarizes the experimental literature on expectations formation. The first column lists the authors and the date of publication. Column (2) displays the number of participants. Column (3) shows the number of process realizations shown at the beginning of the experiment. Column (4) reports the number of rounds of forecasts each participant has to make. Column (5) describes the process. Most of the time, it is an AR(1). In one case, it is an exponentially growing process. In another case, it is an integrated moving average. Column (6) reports if all participants see the same draw or different draws of the same process. Column (7) is "Yes" if the data presented is presented as economic data or not. Column (8) shows that nearly all experiments feature some form of monetary incentives. Column (9) describes the format used to present the data (graphical or number list). Column (10) shows the forecast horizon requested. The last column describes the models tested.

(1) Paper	(2) # of parti-cipants	(3) # of history	(4) # of predic-ctions	(5) Process	(6) Same draw	(7) Econ. back-ground	(8) Monetary Incentives	(9) Format graph /text	(10) Forecast Horizon	(11) Model Tested
Schmalensee (1976)	23	26	27	$\rho \approx 1$	Yes	Yes	Yes	Both	1-5	Adaptive +Extrap.
Andreassen (1990)	77	5	5	$e^{\alpha t}$	Yes	Yes	No	Text	1	Extrap.
DeBondt (1993)	27	48	2	$\rho \approx 1$ non-rep.	Yes	Yes	Weak	Graph	7,13	Extrap.
Dwyer&al (1993)	70	30	40	$\rho = 1$	No	No	Yes	Both	1	Adaptive
Hey (1994)	50	20	48	$\rho \in \{.1, .5, .8, .9\}$	Yes	No	Yes	Both	1	Adaptive
Bloomfield &Hales(2002)	38	9	1	$\rho \approx 1$ non-rep	No	Yes	Yes	Both	1	Extrap.
Asparouhova et al(2009)	92	100	100	$\rho \approx 1$	Yes	No	Yes	Graph	1	BSV vs Rabin
Reimers &Harvey (2011)	2,434	50	Varies	$\rho \in \{0, 0.4, 0.8\}$	No	No	Yes	Graph	1	N/A
Beshears et al(2013)	98	100k	60	ARIMA (0,1,50)	No	No	Yes	Graph	1	Natural Expec.
Frydman &Nave(2016)	38	10	400	$\rho \approx 1$	No	Yes	Yes	Graph	1	Extrap.
This paper	1,600+	40	40	$\rho \in \{0, .2, .4, .6, .8, 1\}$	Both	Both	Yes	Both	1,2,5	Multiple

Table A.2: Summary of Conditions

This table provides a summary of the experiments we conducted. Each panel describes one experiment, and each line within a panel corresponds to one treatment condition. Columns (1) to (3) show the parameters of the AR(1) process $x_{t+1} = \mu + \rho x_t + \epsilon_{t+1}$. Participants are only allowed to participate once.

#	Short description	(1)	(2) AR(1) process	(3)	(4) Number of participants
		persistence ρ	constant μ	volatility σ_ϵ	
Panel A: Experiment 1 – Baseline, MTurk					
A1	Baseline $\rho = 0$	0	0	20	32
A2	Baseline $\rho = 0.2$	0.2	0	20	32
A3	Baseline $\rho = 0.4$	0.4	0	20	36
A4	Baseline $\rho = 0.6$	0.6	0	20	39
A5	Baseline $\rho = 0.8$	0.8	0	20	28
A6	Baseline $\rho = 1$	1	0	20	40
Panel B: Experiment 2 – Long horizon, MTurk					
C1	Horizon: F1 + F5	0.2	0	20	41
C2	Horizon: F1 + F5	0.4	0	20	26
C3	Horizon: F1 + F5	0.6	0	20	31
C4	Horizon: F1 + F5	0.8	0	20	30
Panel C: Experiment 3 – DGP information, MIT EECS					
D1	Baseline	0.2	0	20	42
D2	Baseline	0.6	0	20	52
D3	Display DGP is AR(1)	0.2	0	20	70
D4	Display DGP is AR(1)	0.6	0	20	40

Table A.3: Summary Statistics

Panel A describes demographics of participants. Columns (1) and (2) provide information for participants in Experiment 1 (Baseline, MTurk); columns (3) and (4) for Experiment 2 (Long horizon, MTurk); columns (5) and (6) for Experiment 3 (Describe DGP, MIT EECS). Panel B reports basic experimental statistics, including the total score, the total bonus (incentive payments) paid in US dollars, the overall time taken to complete the experiment, and the time taken to complete the forecasting part (the main part).

Panel A. Participant Demographics

		(1)	(2)	(3)	(4)	(5)	(6)
		Experiment 1		Experiement 2		Experiment 3	
		Obs.	%	Obs.	%	Obs.	%
Gender	Male	117	56.5	67	52.3	88	43.1
	Female	90	43.5	61	47.7	116	56.9
Age	<= 25	30	14.5	18	14.1	197	96.6
	25-45	138	66.7	89	69.5	7	3.4
	45-65	35	16.9	20	15.6	0	.0
	65+	4	1.9	1	.8	0	.0
Education	Grad school	20	9.7	18	14.1	0	.0
	College	132	63.8	74	57.8	207	100.0
	High school	55	26.6	36	28.1	0	.0
	Below/other	0	.0	0	.0	0	.0
Invest. exper.	Extensive	7	3.4	3	2.3	2	1
	Some	58	28.0	29	22.7	43	21.1
	Limited	71	34.3	56	43.8	138	67.7
	None	71	34.3	40	31.3	21	10.3
Taken stat class	Yes	90	43.5	48	37.5	-	-
	No	117	56.5	80	62.5	-	-

Panel B. Experimental Statistics

	Mean	p25	p50	p75	SD	*N*
Experiment 1						
Total forecast score	2,004	1,690	1,990	2,335	462	207
Bonus ($)	3.34	2.82	3.32	3.89	.77	207
Total time (min)	18.01	10.92	13.11	21.85	11.34	207
Forecast time (min)	6.80	4.54	5.66	7.79	3.53	207
Experiment 2						
Total forecast score	1,843	1,588	1,820	2,138	463	128
Bonus ($)	3.07	2.65	3.04	3.57	.77	128
Total time (min)	15.82	8.74	13.11	19.66	9.80	128
Forecast time (min)	6.70	4.54	6.02	7.58	3.17	128
Experiment 3						
Total forecast score	2,071	1,755	2,046	2,326	430	204
Bonus ($)	8.63	7.31	8.53	9.69	1.79	204
Total time (min)	18.47	7.57	10.02	14.09	37.67	204
Forecast time (min)	8.78	4.03	5.09	7.46	19.72	204

Table A.4: Effect of Knowing the Process

This table reports the implied persistence in Experiment 3 among MIT EECS students. Participants are randomly assigned to $\rho = 0.2$ and $\rho = 0.6$. In addition, half of them are randomly assigned to the baseline control condition (control) where the process is described as a stable random process, while the other half are assigned to the treatment condition where they are told that the process is a fixed and stationary AR(1) process.

	Baseline Condition	Knows AR(1)	Test of difference (p-value)
$\rho = .2$	0.56	0.65	0.14
$\rho = .6$	0.86	0.88	0.71

Table A.5: Estimations of Expectations Models

This table reports estimation of eight expectation formation models. Each model is described by an equation and a parameter, highlighted in bold. Estimations are based on pooled data from all conditions of Experiment 1 (i.e., with $\rho \in \{0, .2, .4, .6, .8, 1\}$). All models except constant gain learning and FIRE (which has no parameter) are estimated using constrained least squares. We cluster standard errors at the individual level. The imperfect memory model is estimated by minimizing, over the decay parameter, the mean squared deviation between predicted and realized forecasts. We then estimate standard errors for this model by block-bootstrapping forecasters. The parameter estimate is reported in the third column, along with standard errors in the fourth column. In the fifth column, we report the mean squared error of each model, as a fraction of the sample variance of forecast. Since forecasts in the $\rho = 1$ condition are mechanically much more variable than the forecasts in the $\rho = 0$ condition, we report here the average of this ratio across conditions. This avoids giving too much weight to the low variance (low ρ) conditions.

Model	Equation	Parameter Estimate	Standard Error	mean MSE / $\text{var} F_t x_{t+1}$
Panel A : Backward-looking models				
Adaptive	$F_t x_{t+1} = \boldsymbol{\delta} F_{t-1} x_t + (1 - \boldsymbol{\delta}) x_t$	.17***	(.04)	.53
Extrapolative	$F_t x_{t+1} = (1 + \boldsymbol{\phi}) x_t - \boldsymbol{\phi} x_{t-1}$	-.07***	(.02)	.56
Panel B : Forward-looking models				
FIRE	$F_t x_{t+1} = E_t x_{t+1}$	-	-	.58
Sticky/noisy information	$F_t x_{t+1} = \boldsymbol{\lambda} F_{t-1} x_{t+1} + (1 - \boldsymbol{\lambda}) E_t x_{t+1}$	.14***	(.04)	.56
Diagnostic	$F_t x_{t+1} = E_t x_{t+1} + \boldsymbol{\theta}(E_t x_{t+1} - E_{t-1} x_{t+1})$	.34***	(.04)	.57
Constant gain learning	Rolling regression at t w/ weights: $w_s^t = \frac{1}{\boldsymbol{\kappa}^{t-s}}$	1.06***	(.01)	.56

Table A.6: Model Fit

This table shows the MSE between ρ_h^s in the model in columns (1), (3), and (5), and the MSE between $F_t x_{t+h}$ implied by the model and $F_t x_{t+h}$ in the data in columns (2), (4), (6). Columns (1) and (2) report results for the 1-period forecast; columns (3) and (4) report results for the 2-period forecast; columns (5) and (6) report results for the 5-period forecast. The adaptive expectations model is: $F_t x_{t+1} = \delta x_t + (1-\delta) F_{t-1} x_t$. The traditional extrapolative expectations model is: $F_t x_{t+1} = x_t + \phi(x_t - x_{t-1})$. The sticky expectations model is: $F_t x_{t+h} = (1-\lambda)\rho^h x_t + \lambda F_{t-1} x_{t+h} + \epsilon_{it,h}$. The diagnostic expectations model is: $F_t x_{t+h} = E_t x_{t+h} + \theta(E_t x_{t+h} - E_{t-1} x_{t+h})$. The constant gain learning model is: $F_t x_{t+h} = \widehat{E}_t x_{t+h} = a_{t,h} + \sum_{k=0}^{k=n} b_{k,h,t} x_{t-k}$.

Forecast horizon	$h=1$		$h=2$		$h=5$	
MSE Type	ρ_h^s (1)	Forecast (2)	ρ_h^s (3)	Forecast (4)	ρ_h^s (5)	Forecast (6)
Current model	0.003	496.1	0.001	719.2	0.001	691.0
Adaptive	0.035	495.7	.	.	.	.
Extrapolative	0.064	527.3	.	.	.	.
Sticy	0.117	556.2	0.140	786.1	0.197	814.6
Diagnostic	0.069	521.2	0.115	758.0	0.177	803.3
Constant gain	0.067	526.8	0.039	749.5	0.033	736.3

C Proofs

C.1 Standard Errors of Error-Revision Coefficient

Proposition 3. Assume a univariate regression of centered variables:

$$y_i = \beta x_i + u_i$$

Then, the standard error of the OLS estimate of β is given by:

$$\boxed{s.d.\left(\widehat{\beta} - \beta\right) \approx \frac{1}{\sqrt{N}}\left(\frac{\text{var}y_i}{\text{var}x_i} - \beta^2\right)^{1/2}}$$

Proof. The OLS estimator of β is given by:

$$\widehat{\beta} = \frac{\frac{1}{N}\sum_i x_i y_i}{\frac{1}{N}\sum_i x_i^2} = \beta + \frac{\frac{1}{N}\sum_i x_i u_i}{\frac{1}{N}\sum_i x_i^2}$$

Hence:

$$\sqrt{N}(\widehat{\beta} - \beta) = \frac{\sqrt{N}\frac{1}{N}\sum_i x_i u_i}{\frac{1}{N}\sum_i x_i^2}$$

By virtue of the CLT, we have:

$$\sqrt{N}\frac{1}{N}\sum_i x_i u_i \to N(0, \text{var}(x_i u_i))$$

while:

$$\frac{1}{N}\sum_i x_i^2 \to \text{var}x_i$$

This ensures that:

$$\sqrt{N}(\widehat{\beta} - \beta) \to N(0, \underbrace{\frac{\text{var}(x_i u_i))}{(\text{var}(x_i))^2}}_{=\frac{\text{var}u_i}{\text{var}x_i}})$$

Note that the asymptotic variance can be rewritten:

$$\begin{aligned}\frac{\text{var}u_i}{\text{var}x_i} &= \frac{\text{var}y_i + \beta^2\text{var}x_i - 2\beta\text{cov}(x_i, y_i)}{\text{var}x_i} \\ &= \frac{\text{var}y_i}{\text{var}x_i} - \beta^2\end{aligned}$$

□

Evidently, this ratio is bigger when the variance of x_i is smaller.
Under the error-revision approach, it can easily be shown that:

$$\frac{\text{var}y_i}{\text{var}x_i} = \frac{(1 + \rho^2\theta^2)}{\rho^2\left((1+\theta)^2 + \theta^2\rho^2\right)} \to +\infty \text{ as } \rho \to 0$$

This makes it clear that the error-revision approach does not work well for small ρ because the right-hand-side variable has a small variance, which makes it hard to estimate λ precisely.

The subjective persistence approach does not have this problem as the variance of the right-hand-side variable is just the variance of the process itself, which is non-zero.

C.2 Lemma 1

Proof. The agent has two decisions: first, she decides what information to retrieve (choose $S_t \subseteq \mathcal{S}_t(x^t)$, and second, she chooses the optimal forecast $F_t x_{t+h}$ given the σ-algebra induced by S_t. We solve this backwards; namely, we characterize the optimal forecast for any choice of S_t, and then solve for the optimal S_t given the optimal forecast that it implies.

It is straightforward to see that with a quadratic loss function the optimal forecast for a given choice of S_t is simply the unbiased expectation of x_{t+h} conditional on S_t. Formally, let $F_t^* x_{t+h}(S_t)$ denote the optimal forecast of the agent under S_t, then

$$F_t^* x_{t+h}(S_t) \equiv \arg\min_{F_t x_{t+h}} \mathbb{E}[(F_t x_{t+h} - x_{t+h})^2 | S_t] \Rightarrow F_t^* x_{t+h}(S_t) = \mathbb{E}[x_{t+h}|S_t]. \tag{C.1}$$

It immediately follows that the firms' loss from an imprecise forecast is the variance of x_{t+h} conditional on S_t

$$\mathbb{E}[(F_t^* x_{t+h}(S_t) - x_{t+h})^2 | S_t] = \text{var}(x_{t+h}|S_t). \tag{C.2}$$

Moreover, we can decompose this variance in terms of uncertainty about the long-run mean and variance of short-run fluctuations:

$$\text{var}(x_{t+h}|S_t) = \text{var}((1-\rho^h)\bar{x} + \rho^h x_t + \sum_{j=1}^{h} \rho^{h-j} \varepsilon_{t+j} | S_t) \tag{C.3}$$

$$= (1-\rho^h)^2 \text{var}(\bar{x}|S_t) + \sigma_\varepsilon^2 \sum_{j=1}^{h} \rho^{2(h-j)} \tag{C.4}$$

where the second line follows from:

1. orthogonality of future innovations to S_t that follows from feasibility ($\varepsilon_{t+j} \perp \mathcal{S}(x^t), \forall j \geq 1$);
2. $\text{var}(x_t|S_t) = 0$ since $x_t \in S_t$ by assumption.

It is important to note that the second term in Equation C.4 is independent of the choice for S_t. We can now rewrite the firms' problem as

$$\min_{S_t} \mathbb{E}[(1-\rho^h)^2 \text{var}(\bar{x}|S_t) + C(S_t) | x_t] \tag{C.5}$$

$$s.t. \{x_t\} \subseteq S_t \subseteq \mathcal{S}(x^t), \tag{C.6}$$

where the expectation $\mathbb{E}[.|x_t]$ is taken conditional on x_t because the choice for what information to retrieve happens after the agent observes the context but before information is retrieved.

The next step in the proof is to show that under the optimal information retrieval, the distribution of $\bar{x}|S_t$ is Gaussian. To prove this, we show that for any arbitrary $S_t \in \mathcal{S}(x^t)$, there exists another $\hat{S}_t \in \mathcal{S}(x^t)$ that (1) induces a Gaussian posterior and (2) yields a lower value for the objective function than S_t. To see this, let $S_t \supseteq \{x_t\}$ be in $\mathcal{S}(x^t)$ and let $\hat{S}_t \supseteq \{x_t\}$ be such that

$$\text{var}(\bar{x}|\hat{S}_t) = \mathbb{E}[\text{var}(\bar{x}|S_t)|x_t].$$

Such an $\hat{S}_t$ exists because $\mathcal{S}(x^t)$ is assumed to contain all possible signals on $\bar{x}_t$ that are feasible, so if an expected variance is attainable under an arbitrary signal, it is also attainable by a Gaussian signal. Since both signals imply the same expected variance, to prove our claim, we only need to show that $C(\hat{S}_t) \leq$

$C(S_t)$. To see this, recall that $C(S_t)$ is monotonically increasing in $\mathbb{I}(S_t, x_{t+h}|x_t)$. Thus,

$$C(\hat{S}_t) \le C(S_t) \Leftrightarrow \mathbb{I}(\hat{S}_t, x_{t+h}|x_t) \le \mathbb{I}(S_t, x_{t+h}|x_t). \tag{C.7}$$

A final observation yields us our desired result: by definition of the mutual information function in terms of entropy,[21]

$$\mathbb{I}(S_t; \bar{x}|x_t) = h(\bar{x}|x_t) - \mathbb{E}[h(\bar{x}|S_t)|x_t]. \tag{C.8}$$

Similarly,

$$\mathbb{I}(\hat{S}_t; \bar{x}|x_t) = h(\bar{x}|x_t) - \mathbb{E}[h(\bar{x}|\hat{S}_t)|x_t]. \tag{C.9}$$

It follows from these two observations that

$$C(\hat{S}_t) \le C(S_t) \Leftrightarrow \mathbb{E}[h(\bar{x}|\hat{S}_t)|x_t] \ge \mathbb{E}[h(\bar{x}|S_t)|x_t]. \tag{C.10}$$

The right hand side of this condition is true by the maximum entropy of Gaussian random variables among random variables with the same variance, with equality holding only if both S_t and $\hat{S}_t$ are Guassian (see for example Cover Thomas and Thomas Joy (1991)).[22] This result implies that $C(\hat{S}_t) \le C(S_t)$. Therefore, for any arbitrary $S_t \subset \mathcal{S}_t(x^t)$ such that $\bar{x}|S_t$ is non-Gaussian, we have shown that there exists $\hat{S}_t \subset \mathcal{S}_t(x^t)$ that is (1) feasible and (2) strictly preferred to S_t and (3) $\bar{x}|\hat{S}_t$ is Gaussian.

Hence, without loss of generality, we can assume that under the optimal retrieval of information, $\bar{x}|S_t$ is normally distributed. Now, for a Gaussian $\{x_t\} \subset S_t \subset \mathcal{S}_t(x^t)$, since entropy of Gaussian random variables are linear in the log of their variance, we have:

$$\mathbb{I}(\bar{x}; S_t|x_t) = h(\bar{x}|x_t) - h(\bar{x}|S_t) \tag{C.11}$$

$$= \frac{1}{2}\log_2(\text{var}(\bar{x}|x_t)) - \frac{1}{2}\log_2(\text{var}(x_t|S_t)). \tag{C.12}$$

For simplicity let us define $\tau(S_t) \equiv \text{var}(\bar{x}|S_t)^{-1}$ as the precision of belief about $\bar{x}$ generated by S_t and $\underline{\tau} \equiv \text{var}(\bar{x}|x_t)^{-1}$ as the precision of the prior belief of the agent about $\bar{x}$. It follows that

$$\mathbb{I}(\bar{x}; S_t|x_t) = \frac{1}{2\ln(2)}\ln\left(\frac{\tau(S_t)}{\underline{\tau}}\right), \tag{C.13}$$

$$C(S_t) = \omega\frac{\exp(2\ln(2)\cdot\gamma\cdot\mathbb{I}(\bar{x}; S_t|x_t)) - 1}{\gamma} \tag{C.14}$$

$$= \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^\gamma - 1}{\gamma}. \tag{C.15}$$

[21]For random variables (X, Y), $\mathbb{I}(X; Y) = h(X) - \mathbb{E}^Y[h(X|Y)]$ where for any random variable Z with PDF $f_Z(z)$, $h(Z)$ is the entropy of Z defined as the expectation of negative log of its PDF: $h(Z) = -\mathbb{E}^Z[\log_2(f_Z(Z))]$.

[22]For completeness, here is a brief outline of the proof for maximum entropy of Gaussian random variables. The claim is: among all the random variables X variance σ^2, X has the highest entropy if it is normally distributed. The proof follows from optimizing over the PDF of the distribution of X:

$$\max_{\{f(x)\ge 0: x\in\mathbb{R}\}} -\int_{x\in\mathbb{R}} f(x)\log(f(x))dx \qquad \text{(maximum entropy)}$$

$$s.t. \int_{x\in\mathbb{R}} x^2 f(x)dx - (\int_{x\in\mathbb{R}} x f(x)dx)^2 dx = \sigma^2 \qquad \text{(constraint on variance)}$$

$$\int_{x\in\mathbb{R}} f(x)dx = 1. \qquad \text{(constraint on } f \text{ being a PDF)}$$

Hence, the agent's problem can be rewritten as

$$\min_{S_t} \mathbb{E}\left[\frac{(1-\rho^h)^2}{\tau(S_t)} + \omega\frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^\gamma - 1}{\gamma}\Big| x_t\right] \tag{C.16}$$

$$s.t. \{x_t\} \subseteq S_t \subseteq \mathcal{S}(x^t). \tag{C.17}$$

Finally, since the objective of the agent only depends on the precision induced by S_t, we can reduce the problem to directly choosing this precision, where the constraint on S_t implies bounds on achievable precision: the precision should be bounded below by $\underline{\tau}$, since the agent knows x_t at the time of information retrieval. Moreover, it has to be bounded above by $\text{var}(\bar{x}|x^t)^{-1}$ which the precision after utilizing *all available information*. Replacing these in the objective, and changing the choice variable to $\tau(S_t)$ we arrive at the exposition delivered in the lemma. □

C.3 Proposition 1

Proof. We start by solving the simplified problem in Lemma 1. The problem has two constraints for τ: $\tau \geq \underline{\tau}$ and $\tau \leq \bar{\tau}(x^t) \equiv var(\mu|x^t)^{-1}$. By assumption $\text{var}(\mu|x^t)$ is arbitrarily small so we can assume that the second constraint does not bind. The K-T conditions with respect to τ are

$$-\frac{(1-\rho^h)^2}{\tau^2} + \frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^\gamma \geq 0, \quad \tau \geq \underline{\tau}, \quad \left(-\frac{(1-\rho^h)^2}{\tau^2} + \frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^\gamma\right)(\tau - \underline{\tau}) = 0.$$

Therefore, the variance of the agent's belief about the long-run mean is given by

$$\text{var}(\mu|S_t) = \tau^{-1} = \underline{\tau}^{-1} \min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{C.18}$$

The next step is to find an optimal signal set $S_t \supseteq \{x_t\}$ that generates this posterior. Two cases arise:

1. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) \geq 1$, then $\sigma^2 = (1-\rho^h)^2\underline{\tau}$ and $S_t = \{x_t\}$ delivers us the agent's posterior. In other words, $\text{var}(\mu|S_t) = \text{var}(\mu|x_t)$ meaning that the agents does not retrieve any further information other than what is implied by the context. In this case, $\mathbb{E}[\mu|S_t] = \mathbb{E}[\mu|x_t] = x_t$ and

$$\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|S_t]|\mu, x_t] = (1-\rho^h)\mathbb{E}[\mathbb{E}[\mu|S_t]|\mu, x_t] + \rho^h\mathbb{E}[\mathbb{E}[x_t|S_t]|\mu, x_t] = x_t \tag{C.19}$$

   and

$$\sigma^2 \equiv var(\mathbb{E}[x_{t+h}|S_t]|\mu, x_t) = var(x_t|\mu, x_t) = 0; \tag{C.20}$$

2. if $\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right) < 1$, then it means that the agent retrieves more information than what is revealed by the context x_t. Suppose a signal structure $\tilde{S}_t$ generates this posterior variance. By Lemma 1 this has to be Guassian. Our claim is that the set $\hat{S}_t \equiv \{x_t, \mathbb{E}[\mu|\tilde{S}_t]\}$ also generates this posterior. Note that elements of this set are also distributed according to a Gaussian distribution. To see the equivalence of the two sets, note that by the law of total variance,

$$\text{var}(\mu|x_t) = \text{var}(\mu|\tilde{S}_t) + \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t)$$
$$\text{var}(\mu|x_t) = \text{var}(\mu|\hat{S}_t) + \text{var}(\mathbb{E}[\mu|\hat{S}_t]|x_t),$$

   but note that

$$\text{var}(\mathbb{E}[\mu|\hat{S}_t]|x_t) = \text{var}(\mathbb{E}[\mu|x_t, \mathbb{E}[\mu|\tilde{S}_t]]|x_t) = \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t).$$

Thus, it has to be that

$$\text{var}(\mu|\tilde{S}_t) = \text{var}(\mu|\hat{S}_t)$$

and the two sets generate the same posterior variance for the agent. Now, note that by Bayesian updating of Gaussians:

$$\mathbb{E}[\mu|S_t] = \mathbb{E}[\mu|\tilde{S}_t] = \mathbb{E}[\mu|x_t] + \frac{\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t)}{\text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t)}(\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|x_t]).$$

Since $\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|x_t] \neq 0$ almost surely, this implies that

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = \underline{\tau}^{-1} - \tau^{-1}, \tag{C.21}$$

where the last equality follows from the law of total variance. Now, consider the following decomposition of $\mathbb{E}[\mu|\tilde{S}_t]$:

$$\mathbb{E}[\mu|\tilde{S}_t] = a\mu + bx_t + \varepsilon_t,$$

where a and b are constants and ε_t is the residual that is orthogonal to both x_t and μ conditional on $\tilde{S}_t$. We have

$$x_t = \mathbb{E}[\mu|x_t] = \mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|x_t] = a\mathbb{E}[\mu|x_t] + bx_t = (a+b)x_t,$$

so $a + b = 1$. Moreover, we also have

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|x_t) = a\text{var}(\mu|x_t),$$

so $a = 1 - \frac{\underline{\tau}}{\tau}$. Therefore,

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|\mu, x_t] = (1 - \frac{\underline{\tau}}{\tau})\mu + \frac{\underline{\tau}}{\tau}x_t \\ \Rightarrow &\mu_t \equiv \mathbb{E}[\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t] = (1-\rho^h)(1 - \frac{\underline{\tau}}{\tau})\mu + (1-\rho^h)\frac{\underline{\tau}}{\tau}x_t + \rho^h x_t. \end{aligned} \tag{C.22}$$

Moreover,

$$\begin{aligned} &\text{var}(\mathbb{E}[\mu|\tilde{S}_t]|x_t) = a^2\text{var}(\mu|x_t) + \text{var}(\varepsilon_t) \\ \Rightarrow &\text{var}(\varepsilon_t) = \frac{1}{\tau}(1 - \frac{\underline{\tau}}{\tau}) \\ \Rightarrow &\sigma^2 \equiv \text{var}(\mathbb{E}[x_{t+h}|\tilde{S}_t]|\mu, x_t) = (1-\rho^h)^2\text{var}(\varepsilon_t) = (1-\rho^h)^2\frac{1}{\tau}(1 - \frac{\underline{\tau}}{\tau}). \end{aligned} \tag{C.23}$$

Plugging in the expression for τ from (C.18) into (C.22) and (C.23) and setting $\mu = 0$ gives us the expressions in the Proposition.

Combining Equations (C.19), (C.20), (C.18), (C.22), (C.23) and setting $\mu = 0$ gives us:

$$\mu_t = \min\left\{1, \rho^h + (1-\rho^h)\left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\} x_t \tag{C.24}$$

$$\sigma^2 = (1-\rho^h)^2\underline{\tau}^{-1}\max\left\{0, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\left(1 - \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right)\right\}. \tag{C.25}$$

□

C.4 Proposition 2

Proof. From Proposition 1 we can derive Δ as

$$\Delta = (1-\rho^h)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{C.26}$$

1. Note that if $\Delta = 0$ then either $\rho^h = 1$ or $\omega = 0$, but recall that this expression for the precision of the long-run mean was derived under the assumption that $\text{var}(\mu|x^t)$ is arbitrarily small. So $\Delta = 0$ if and only if either $\rho = 1$ or $\omega = 0$ and past information potentially available to the forecaster is infinite.
2. As long as $\gamma \geq 0$, which is true by assumption, it is straightforward to verify that Δ is increasing in ω and $\underline{\tau}$.
3. For Δ to be decreasing in ρ^h it has to be the case that $(1-\rho^h)^{1-\frac{2}{1+\gamma}}$ is decreasing in ρ^h, which is the case if and only if

$$1 - \frac{2}{1+\gamma} \geq 0 \Leftrightarrow \gamma \geq 1. \tag{C.27}$$

□

C.5 Corollary 1

Proof. From Proposition 2 we have

$$\ln(\zeta) = \ln\left(1 + (\rho^{-h} - 1)\min\left\{1, \left(\frac{\omega\underline{\tau}}{(1-\rho^h)^2}\right)^{\frac{1}{1+\gamma}}\right\}\right). \tag{C.28}$$

First of all, it is straightforward to see that the term in side the log in the right-hand side is larger than 1, so perceived persistence is larger than actual persistence — in other words, ζ is a measure of overreaction. Moreover, for ζ to be decreasing in ρ^h it has to be the case that $(1-\rho^h)^{1-\frac{2}{1+\gamma}}/\rho^h$ is decreasing in ρ^h, which is true if and only if $\gamma \geq 2\rho^h - 1$. Therefore for ζ to be decreasing for any value of ρ^h, it has to be the case that $\gamma \geq 1$. □