Abstract

Experimental studies have shown different degrees of randomness both over time and across different setups. This paper proposes Thompson sampling as a way to explain this randomness endogenously. Thompson sampling means that individuals update their subjective probability distribution of unknown parameters in a Bayesian way, make a random draw from the posterior, and then act optimally, conditional on that draw. For different experimental datasets (2x2 games and forecasting in markets), the empirical fit of Thompson sampling is compared to two other hypotheses of noisy decision-making: Bayesian learning with exogenous shocks and quantal response equilibrium (QRE). In datasets where the amount of randomness does not vary, the difference between Thompson sampling and other models is insignificant. Conversely, in datasets where the amount of randomness varies substantially, Thompson sampling provides a better fit than the other two benchmarks.

Keywords: Learning, adaptive learning, Bayesian learning, behavioral game theory, expectations, stochastic choice

JEL Classifications: C91, C92, D84, E37
1 Introduction

Numerous studies in economics find that individual choices can be considered as stochastic: when individuals are asked to make repeated choices from the same set of options, they are frequently observed to make different choices over time. Randomness in choice has often been interpreted as driven by unobservable factors. (See e.g. Train (2009)) Even in an artificial laboratory setup, in which the experimenter has full information and control over observables, there is frequently a myriad of unobservables stemming from cognitive factors. Due to the presence of those unobservables, an agent’s choice cannot be predicted exactly and thus contains a random element.

Traditionally, randomness has been introduced into many areas of economics by exogenous shocks. Not only has this been common practice in macroeconomics and econometrics, but also in microeconomics, where exogenous shocks are used for example by random utility models (McFadden 1974; Gul and Pesendorfer 2006) and noisy generalizations of equilibrium such as quantal response equilibrium (McKelvey and Palfrey 1995).

Yet, those exogenous shocks are arbitrary and thus cannot explain why noise patterns differ systematically across setups and time. Many experimental studies document that the amount of noise is greater in some environments than others (e.g. Hommes et al. (2005), Fehr and Tyran (2008), Heemijer et al. (2009), Mauersberger (2016)) and that the amount of noise observed varies over time (e.g. Nagel (1995)). Applications of quantal response equilibrium to laboratory data have thus frequently demonstrated structural breaks over time in their exogenous parameters (see e.g. McKelvey and Palfrey (1995)) or have allowed their exogenous parameters to differ across different experiments (see e.g Dufwenberg et al. (2007)). The fact that models like quantal response do not endogenously account for these changing noise patterns inhibits their use for predictive purposes. This paper thus introduces Thompson sampling (Thompson, 1933) as a different way of looking at stochastic decision-making in situations of strategic interaction in economics.

Consider penalty kicks in soccer as an illustrative example for decision-making. The player has to decide whether to kick the ball into the left-hand side or into the right-hand side of the goal, and the goalkeeper has to decide where jump. While there is a unique mixed strategy...

---

1Randomness in choice has not only been documented in studies of individual decision-making but also in strategic interaction. For randomness in individual decision-making refer e.g. to Mosteller and Nogee (1951); Hey (1995); Ballinger and Wilcox (1997); Cheremukhin et al. (2011). For noise in strategic interaction, see e.g. McKelvey and Palfrey (1995); Roth and Erev (1998); Camerer and Ho (1999).
Nash equilibrium of playing both left and right with 50% probability, the actual probability with which the player plays left or right may in practice be less clear due to for example different player abilities. Thus, the goalkeeper has to gauge or estimate these probabilities. Assume the goalkeeper uses Bayesian inference. Then she first has some initial belief, called prior distribution. Suppose the goalkeeper has watched the specific player in many previous matches, then those observations are used to update the prior distribution using Bayes’ rule. The resulting distribution is the posterior distribution. If the goalkeeper is an expected utility maximizer, she would always use a well-defined point of the posterior to make her decision: the expected value. For example, if the expected value of the player choosing left is 50.1%, the goalkeeper would always jump to the left.

If she nevertheless jumped to the right, it would represent an exogenous shock in the sense that the classical expected utility model could not explain that.

Thompson sampling departs from the view that the decision-maker always uses the expected value. Instead, it postulates that the decision-maker makes a random draw from the posterior. Conditional on that draw, the agent acts optimally. If the goalkeeper in the penalty example is a Thompson sampler, she might end up drawing a 60% probability of the player choosing left, which would urge her to jump to the left; alternatively, she could draw a 45% probability of the player choosing left from the posterior, which would urge her to jump to the right.

This has two important implications for randomness. First, shocks are endogenous in the sense that the distribution from which the element is drawn changes over time due to Bayesian learning. Second, randomness is introduced into belief formation rather into the selection of an action.

Thompson sampling has been widely used in psychology and computer science, but has, to the best of my knowledge, not previously been used in economics. I show that Thompson sampling is widely applicable over many setups in economics, not only comprising simple microeconomic games on the discrete action space but also setups in macroeconomics. Hence, Thompson sampling could be considered a unifying approach to the literature of learning and out-of-equilibrium decision-making.

---

3 Agranov and Ortoleva (2017) show in an experimental study that the majority of subjects has an explicit preference for randomization.

2 This paper postulates a hypothesis about endogenous noise, not about how people update. While different ways of updating from Bayesian are thinkable, I abstract from this question by assuming that people update optimally and hence in a Bayesian way. Moreover, Thompson sampling is defined as using Bayesian updating.

4 For a psychological study in which the fit of Thompson sampling has been compared to other Bayesian and non-Bayesian algorithms in experimental bandit tasks, see for instance Speekenbrink and Konstantinidis (2015). For the normative literature about Thompson sampling in bandit tasks in computer science, refer for example to May et al. (2012); Agrawal and Goyal (2012a,b); Kaufmann et al. (2012); Scott (2010); Chapelle and Li (2011).
making, which has been quite disconnected across the two subfields in economics, microeconomics and macroeconomics.

This paper investigates to what extent Thompson sampling has descriptive power in different setups of interactive decision-making. As benchmarks for comparison, I use two models of noisy decision-making: quantal response equilibrium (QRE) (McKelvey and Palfrey, 1995) and Bayesian learning with exogenous shocks from a type I extreme value distribution, so that actions are chosen with the probabilities of a logit model.

I apply those three models to laboratory data, because they represent controlled environments in which most confounding factors that could drive randomness are eliminated and the cognitively-driven randomness purported by Thompson sampling can be observed. Second, the information given to subjects is controlled so that there is little ambiguity of which observations the subjects use to update their beliefs.

The datasets used in this paper represent very different setups: The first dataset is taken from Erev et al. (2007) and contains multiple rounds of ten different 2x2 games with perfect and complete information. The ten games are distinct by different payoff matrices and observed randomness in play is greater in some games than in others. The second dataset contains learning-to-forecast experiments by Heemeijer et al. (2009), in which agents are asked to forecast prices in two different markets: one in which forecasting decisions are strategic substitutes and one in which forecasting decisions are strategic complements. In the market with strategic substitutes, dynamics quickly converge to the fundamental and noise tends to fade, while in the market with strategic complements, dynamics do not converge to the equilibrium and display persistent fluctuations.

Thompson sampling is particularly successful in endogenously predicting different dynamics across environments such as the differences between strategic substitutes and complements in the learning-to-forecast experiments. In settings such as the 2x2 games, where differences in behavior caused by unobservable factors are less distinct, there is no evidence that Thompson sampling provides a better fit than other models.

---

5Data from soccer penalty kicks have been used before in the economics literature (see e.g. Apesteguia and Palacios-Huerta (2010)). Yet, they have the limitation that they do not contain long time series, as they only contain few observations of every goalkeeper-player pair.
2 General model

I take the version of Thompson sampling outlined by Chapelle and Li (2011) and modify it for interactive games.

Setup Assume there is a finite set of players $I = \{1, 2, ..., n\}$. A given player is referred to as $i \in I$, while this player’s opponents are referred to as $-i$. The game has a set of rounds $\mathcal{T}$, which may either be finite so that $\mathcal{T} = (1, 2, ..., T)$ or infinite so that the game is infinitely repeated. A particular round from this set is referred to as $t \in \mathcal{T}$.

Player $i$ chooses a specific action $a_i^t \in A^i$ out of a set of actions $A^i$, which can be either a continuous or a discrete action space. The probability of player $i$’s play in period $t$ is given by a probability distribution $\sigma_i^t \in \Delta$, where $\Delta$ is the set of all possible probability distributions. $\sigma_i^t$ is a probability density if $A^i$ is continuous and a probability mass function if $A^i$ is discrete.

The individual history of actions of player $i$ at period $t$ is defined to be $i$’s sequence of actions from period 1 to period $t-1$: $h_i^t \equiv (a_i^1, a_i^2, ..., a_i^{t-1}) \in \mathcal{H}_i^t$, where $\mathcal{H}_i^t$ is the set of histories at time $t$ for player $i$. The collective history of actions for all players is then $h_t \equiv \bigcup_{i=1}^{|I|} \{h_i^t\} = \{(a_1^1, ..., a_{i-1}^1), (a_1^2, ..., a_{i-1}^2), ..., (a_{n}^1, ..., a_{n}^{t-1})\} \in \mathcal{H}_t$, where $\mathcal{H}_t$ is the set of collective histories at time $t$.

Assume there are $m$ payoff-relevant states of the world $\omega_t \equiv \{\omega_1^t, \omega_2^t, ..., \omega_m^t\} \in \Omega$, where $\Omega$ denotes the set of all possible realizations of these states of the world. For every player $i$ at every period $t$, there is a (stochastic) payoff mapping

$$u_i^t : A^i \times \mathcal{H}_i^t \times \Omega \times A^{-i} \to \mathbb{R}$$

(1)
giving von Neumann-Morgenstern utility $u_i^t$ to each player $i$ in every period $t$. Bayesian updating comes into play, if there is some uncertainty.

Assume that every player $i$ has some kind of uncertainty, so that player $i$ has to form beliefs about some of the variables. Denote $\Gamma^i$ the set of all possible realizations of those unknown variables. For example, if player $i$ at time $t$ does not know $a_{-i}^t$ and $\omega_t$, then $\Gamma^i = A^i \times \Omega$. Similarly, let $K^i$ be the set of all possible realizations of the known variables. Let $k_i^t \in K^i$ be a vector realization of the known variables for player $i$ at time $t$. Similarly, let $\gamma_i^t \in \Gamma^i$ be a vector realization of the unknown variables for player $i$ at time $t$. 


Then the expected utility can be written as

$$E_t u_i^t : K^i \times \Delta(\Gamma^i)$$

(2)

where $\Delta(\Gamma^i)$ denotes the set of all joint distributions using the unknowns as an input.

A player $i$’s information about $\gamma_i^t$ can be summarized by a profile of (stochastic) signal functions $z_i^t : \Gamma^i \rightarrow Z^i$, where $Z^i$ is the space of all possible signal realizations for player $i$. This implies an objective likelihood of observing a particular signal $z_i^t(\gamma_i^t)$, which can be denoted by $p^i(z_i^*(\gamma_i^t)|\cdot)$.

Without loss of generality, assume that a signal merely becomes available at the end of a period $t$. Then the individual history of signals can be denoted by $\zeta_i^t \equiv (z_i^1(\gamma_i^1), z_i^2(\gamma_i^2), ..., z_i^{t-1}(\gamma_i^{t-1}))$.

Importantly, I allow for the possibility that a player $i$ misperceives the likelihood of the signal so she believes in a subjective likelihood being given by $p^i(z_i^t(\gamma_i^t)|\theta_i)$, where $\theta_i$ is a vector of unknown parameters associated with the subjective likelihood.

**Bayesian updating** Thompson sampling assumes Bayesian updating. Let agent $i$’s subjective prior belief about the vector of unknown parameters at time $t$, $\theta_i$, be the posterior conditional upon the history of signals at time $t$: $\mu_i(\theta_i|\zeta_i^t) \in \Delta$. After each player $i$ has chosen her action $a_i^t$ and has observed the signal realizations $z_i^t$, the posterior distribution becomes

$$\mu_i(\theta_i|\zeta_i^{t+1}) \propto p^i(z_i^t(\gamma_i^t)|\theta_i) \cdot \mu_i(\theta_i|\zeta_i^t)$$

(3)

**Choosing an action** The novelty that is introduced by Thompson sampling is how beliefs are formed. I assume that agents choose their action myopically, i.e. agents’ sole objective is maximizing the expected payoffs of period $t$ so that an agent does not explicitly take into account how her action affects the set of her future payoffs $(u_{t+1}^i, ..., u_T^i)$. Then, the agent’s objective is choosing the optimal action $a_i^t$ so that:

$$a_i^t = \arg \max_{a_i^t} E(u_i^t(k_i^t, \gamma_i^t))$$

(4)

---

6In the K-armed bandit problem with $n = 1$, $A$ corresponds to a discrete set of several bandits from which the agent needs to choose. In this case there are $K$ unknown parameters $\omega_1, \omega_2, ..., \omega_k, ..., \omega_K$, where $\omega_k$ corresponds to the reward generated in arm $k$. Since the bandit setting has been well-addressed in the psychology literature, this paper does not elucidate this setting further.
Pure belief learning models assume that players would choose $a^i_t$ by computing the expected value of $u^i_t$ (assuming that it exists) over the entire space of $\theta^i$: 

$$a^i_t = \arg \max_{a^i_t} \mathbb{E}(u^i_t(k^i_t, \gamma^i_t)) = \int \int u^i_t(k^i_t, \gamma^i_t) \cdot p^i(\gamma^i_t(\gamma^i_t) | \theta^i) \cdot \mu^i(\theta^i | \zeta^i_t) \partial \gamma^i_t \partial \theta^i$$  \hspace{1cm} (5)

Since the expected value is a fixed point for any history of signals $\zeta^i_t$, this results in a theory of deterministic play.

Thompson sampling departs from that approach and introduces a theory of stochastic beliefs. Similarly to the bandit literature, as $\theta^i$ is unknown, the individual has a dual objective: acquiring new information (exploration) and payoff maximization (exploitation.) Thompson sampling belongs to a class of heuristics referred to as probability matching, in which the decision-maker randomly selects an action $\tilde{a}^i_t$ according to its probability of maximizing the payoff conditional on her beliefs:

$$p(a^i_t = \tilde{a}^i_t) = p^i\{\tilde{a}^i_t = \arg \max_{a^i_t} \mathbb{E}(u^i_t(k^i_t, \gamma^i_t))\}$$  \hspace{1cm} (6)

This probability must be computed or estimated due to the unknown parameter $\theta^i$.

Within that class of heuristics, Thompson sampling is the specific case, in which $\theta^i$ is updated in a Bayesian way. Hence, the probability of a specific action $\tilde{a}^i_t$ being optimal to play in period $t$ depends on the posterior of the parameter $\theta^i$ given the history of actions and payoffs until period $t-1$:

$$p(a^i_t = \tilde{a}^i_t) = \int \mathbb{I}[\tilde{a}^i_t = \arg \max_{k^i_t} \mathbb{E}(u^i_t(k^i_t, \gamma^i_t)) | \mu^i(\theta^i | \zeta^i_t)] \mu^i(\theta^i | \zeta^i_t) \partial \theta^i$$  \hspace{1cm} (7)

where $\mathbb{I}$ is the indicator function, being 1 if the expression in square brackets is true and 0 otherwise. However, (7) does not have to be computed explicitly. The objective in (6) is attained by making a random draw $\tilde{\theta}^i_t$ from its probability distribution $\mu^i(\theta^i | \zeta^i_t)$ and acting optimally conditional on this draw.

In summary, the Thompson sampling algorithm works in the following way:

In each period $t = 1, ..., T$, every player $i$ proceeds the following way:

1. Draw a realization $\tilde{\theta}^i_t$ from the posterior distribution $\mu^i(\theta^i | \zeta^i_t)$

2. Choose $a^i_t = \arg \max_{a^i_t} \mathbb{E}(u^i_t(k^i_t, \gamma^i_t))$
3. Observe the payoff $u_i^t$.

### 2.1 Relation to the previous literature

The learning and decision theory literature has proposed other specific hypotheses of stochastic play, which frequently use a random utility approach \cite{McFadden1974}. That class of models assumes a random element in a utility mapping: $U : A_i^t \rightarrow \mathbb{R}$. The utility for any arbitrary action $\tilde{a}_i^t$ can be decomposed into:

$$U_i^t(a_i^t = \tilde{a}_i^t) = V(a_i^t = \tilde{a}_i^t) + \epsilon_i^t$$

where $V(.)$ determines the deterministic part, $\epsilon_i^t(.)$ represents a random element. The distribution of this random element determines the probability distribution of the actions $Pr(a_i^t = \tilde{a}_i^t)$. A common assumption is a type I extreme value distribution for the errors with variance $\pi^2/6$, since this results in a convenient multinomial logit specification for the actions \cite{Train2009} (See e.g. \cite{Train2009} for details.) As there is typically no good reason why the error variance should be $\pi^2/6$, this requires normalizing the utility specification in (8). Denote $Var(\epsilon_i^t) \equiv \sigma^2 : \pi^2/6$ and suppose, without loss of generality, there is an action set $A_i^t = \{a_i^{i1}, ..., a_i^{iJ_i}\}$, where $J_i$ is the number of pure strategies.

Then, a multinomial logit specification is given by:

$$Pr(a_i^t = a_i^{ij}) = \frac{\exp(\frac{1}{\sigma}V(a_i^t = a_i^{ij}))}{\sum_{j=1}^{J_i}\exp(\frac{1}{\sigma}V(a_i^t = a_i^{ij}))}$$

(9)

This specification, being often referred to as softmax learning, is widely used in the learning literature. Redefining $\lambda \equiv \frac{1}{\sigma}$ yields an exogenous parameter, governing the rationality of agents. If $\sigma \rightarrow \infty$ or $\lambda = 0$, agents can be considered completely irrational, since they would choose each action with equal probability. Conversely, as $\sigma = 0$ or $\lambda \rightarrow \infty$ players become completely rational.

Since $\lambda$ is exogenous, the amount of randomness in these logit-type models is by design also exogenous. This is a stark assumption, which is also at odds with many experimental studies (see introduction). While applications of the logit quantal response equilibrium allow for different

\footnote{An alternative specification in the previous learning literature is a power probability specification of which the linear Luce specification used in early reinforcement learning literature is a special case. For a discussion of the relative merits and drawbacks of the power and the logit specifications, see \cite{CamererHolt1999}. Less common is a multinomial probit specification due to computational difficulties. (See \cite{CameronTrivedi2005} for a discussion.)}
\(\lambda\)-parameters over settings (Dufwenberg et al., 2007) and over time (McKelvey and Palfrey, 1995), the exogeneity of this parameter impedes the use of those models for predictive purposes.

Instead of focusing on the random part \(\epsilon_t\), the previous learning literature has focused on the specification of the deterministic part of the random utility function \(V(a_i^t = \tilde{a}^i)\). Quantal response equilibrium (QRE) (McKelvey and Palfrey, 1995) is a specific hypothesis, purporting that this part is calculated rationally in the sense that \(V(.)\) is the expected value of the payoff, given equilibrium beliefs about other players' actions. Other approaches, including reinforcement learning (Roth and Erev, 1998), heuristic-switching (Brock and Hommes, 1997) and experienced-weighted attraction learning (Camerer and Ho, 1999), assume a boundedly rational way of calculating \(V(.)\), not using the assumption of equilibrium beliefs.

<table>
<thead>
<tr>
<th>Beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equilibrium</td>
</tr>
<tr>
<td>Optimal</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Exogenous</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Optimal response to sample from posterior</td>
</tr>
</tbody>
</table>

Table 1: Thompson sampling vs. previous approaches

While most previous approaches assume random perturbations, randomness evolves endogenously for Thompson sampling, because it is governed by the Bayesian posterior, summarizing information acquired about the environment. Specifically, Thompson sampling deviates from previous approaches in the literature as summarized in table 1. The contrast to Nash equilibrium, perfect foresight equilibrium and rational expectations (first row left) is twofold: first, agents using Thompson sampling do not form equilibrium beliefs but instead Bayesian beliefs. The second dimension is how agents select their actions. While Thompson sampling defines the distribution of the random elements endogenously as the Bayesian posterior, randomness under equilibrium beliefs does not occur other than due to deterministic (and optimal) responses to exogenous randomness in fundamentals. As stressed previously by Roth and Erev (1998), pure belief learning models like fictitious play (first row center) correspond to Bayesian updating with deterministic action selection. Another example in this class of models is the “internal rationality” approach
by Adam and Marcet (2011). Bayesian learning can be augmented by introducing an exogenous random component determining agents’ choice (second row center). Other models postulate non-Bayesian learning hypotheses (second row right) and commonly add exogenous trembles to create probabilistic choice. Examples include reinforcement learning (Roth and Erev, 1998), experience-weighted attraction (Camerer and Ho, 1999) and heuristic-switching models (Brock and Hommes, 1997).

In order to evaluate Thompson sampling empirically, one needs a benchmark model against which one could potentially reject Thompson sampling. I use two criteria for choosing this benchmark model: first, since Thompson sampling is widely applicable, the benchmark model should also be widely applicable. One could certainly test Thompson sampling against many specific models in many specific contexts. However, if a model has been developed for a specific context, it would be less surprising if that specific model provided a better fit for the context it has been developed for than Thompson sampling. Second, just as Thompson sampling, the benchmark model should be well-defined and not be arbitrary. In fact, once one departs from calculating $V(.)$ rationally, there is a myriad of possible specifications. Even though models like reinforcement learning and experience-weighted attraction have foundations in psychology, many different versions of them are thinkable and have been used in the previous literature. I focus on the comparison between Thompson sampling and QRE: QRE has been widely applied for all kinds of games as well as in macroeconomics (Costain and Nakov, 2015), and uses a non-arbitrary belief structure by assuming consistent beliefs. Thompson sampling has the advantage that it is easier to implement, as QRE requires an (often intractable) fixed point calculation.

Table 1 highlights that quantal response equilibrium differs from Thompson sampling in two vital dimensions: while Thompson sampling is a non-equilibrium concept, QRE is an equilibrium concept. The second dimension is the error structure. While the distribution of the errors is arbitrary, it is usually taken as logit for QRE. (See Camerer and Ho (1999) for a discussion of different error structures.) However, logit is not a convenient specification for Bayesian learning due to intractability. (See e.g. Koop and Poirier (1993).) To disentangle whether any difference between Thompson sampling and QRE comes from the way beliefs are specified or from the error structure, I also provide a step in between with Bayesian learning in the same way as specified

---

8The left cell of the third row notes that drawing shocks from the posterior and being in equilibrium implies a contradiction: in order to yield a posterior, agents need to engage in Bayesian updating, which implies non-equilibrium behavior. The right cell of the third row notes that sampling from the Bayesian posterior is not applicable if the belief formation process is different from Bayesian.
for Thompson sampling and a logit error structure, labeled as “Logit” (second row center). The approach labeled as “logit” can be considered to be a hybrid between Thompson sampling and QRE, since it uses the logistic error structure of QRE but an estimate for the expected payoff using the same belief specification as in Thompson sampling. Below I briefly review quantal response equilibrium for the unfamiliar reader and provide an exposition for the Bayesian logit approach.

**Quantal response equilibrium (QRE)** McKelvey and Palfrey (1995) assume that the deterministic part of the random utility function is the expected payoff. Calculation of the expected payoff requires specifying subjective beliefs of every player about the distribution of other players’ actions $a_{it}^{\neq i}$. McKelvey and Palfrey (1995) assume that these subjective beliefs are consistent with the actual probability distribution of other players under the quantal response equilibrium hypothesis. Hence, the expected payoff is only conditioned on a player’s own action and the probability distribution of others $Pr(a_{it}^{\neq i})$

$$V(a_i^t = a_{ij}) = E(u_i^t|a_i^t = a_{ij}, Pr(a_{it}^{\neq i}))$$ (10)

Assuming a type I extreme value distribution, the probability distributions over the actions are given by the logit model:

$$P(a_i^t = a_{ij}) = P(U_i^t(a_i^t = a_{ij}) \geq U_i^t(a_i^t = a_{ik}), \forall k \in J_i) = \frac{\exp(\lambda E(u_i^t|a_i^t = a_{ij}, Pr(a_{it}^{\neq i}))}{\sum_{k=1}^{J_i} \exp(\lambda E(u_i^t|a_i^t = a_{ik}, Pr(a_{it}^{\neq i}))}$$ (11)

**Bayesian logit** While quantal response equilibrium (QRE) is an equilibrium concept in which subjective beliefs coincide with objective beliefs, this assumption can be relaxed in favor of non-equilibrium beliefs about other players’ probability of play $\mu^i(a_{it}^{\neq i})$ or a perceived law of motion (PLM) about how the (market) outcomes are generated that may not coincide with the actual law of motion (ALM)\(^{10}\). Those non-equilibrium beliefs $\mu^i(a_{it}^{\neq i})$ may contain states that are unknown to the decision-maker and which she has to learn. It is assumed that as new information is available, the decision-maker updates $\mu^i(a_{it}^{\neq i})$ in a Bayesian way. Given her Bayesian beliefs, she calculates

---

\(^9\)As every player $i$ forms consistent beliefs and rationally calculates the reward, the past does not play any role.

\(^{10}\)PLM and ALM is the terminology used by the macroeconomic learning literature. See for example Evans and Honkapohja (2001).
the expected reward

\[ V(a_i^t = a^{ij}) = \mathbb{E}(u_i^t | a_i^t = a^{ij}, \mu^i(a_{i-1}^{−i})) \] (12)

If one assumes a type I extreme value distribution similarly to the application of quantal response equilibrium above, the probability distribution over the action space is given by the logit model:

\[ P(a_i^t = a^{ij}) = \frac{\exp(\lambda \mathbb{E}(u_i^t | a_i^t = a^{ij}, \mu^i(a_{i-1}^{−i}))}{\sum_{k=1}^{J} \exp(\lambda \mathbb{E}(u_i^{t+1} | a_i^t = a^{ik}, \mu^i(a_{i-1}^{−i})))} \] (13)

3 Application to 2x2 games

3.1 Theory

There are \( n = 2 \) players, row (ROW) and column (COL). The set of actions for both players is \( A^i \) with two actions \( a^{i,1}, a^{i,2} \). (For example, \( a^{ROW,1} = T, a^{ROW,2} = B, a^{COL,1} = L, a^{COL,2} = R \).) The game is repeated \( \tau \) rounds, indexed by \( t = 1, 2, ... \tau \). The probability distributions of play are \( \sigma^*_t = (p^i_t(a^{i,1}), p^i_t(a^{i,2})) = (p^i_t(a^{i,1}), 1 - p^i_t(a^{i,2})) \), with \( p^i_t(a^{i,1}) \) denoting the objective probability that player i plays \( a^{i,1} \) in period t. The payoff \( u_i^t \) is given by a von-Neumann-Morgenstern payoff matrix in table 2.

<table>
<thead>
<tr>
<th>ROW</th>
<th>( a^{ROW,1} )</th>
<th>( a^{ROW,2} )</th>
<th>( a^{COL,1} )</th>
<th>( a^{COL,2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( P_1^{ROW} ), ( P_1^{COL} )</td>
<td>( P_2^{ROW} ), ( P_2^{COL} )</td>
<td>( P_3^{ROW} ), ( P_3^{COL} )</td>
<td>( P_4^{ROW} ), ( P_4^{COL} )</td>
</tr>
</tbody>
</table>

Table 2: Exemplary payoff matrix

Under perfect and complete information, any player i’s only unknown at time t is the opponent’s contemporaneous action \( a_{i−1}^{−i} \). However, she observes all past actions directly so that they constitute the observed signals \( z_{i−1}^{−i} = a_{i−1}^{−i} \), so that \( \zeta^*_t = (a_{1−1}^{−i}, ..., a_{t−1}^{−i}) \).

As it is relevant for the subsequent applications, the payoffs could also represent lotteries: for example, if both players play L, ROW receives a fixed payoff \( \nu \) with probability \( P_1^{R} \), while COL does with probability \( P_1^{C} \).
Belief formation about the other player’s action  Player i does not know the likelihood of her opponent’s play \( p^*(a_i^{-i}|.) = \sigma^i_t \). Hence, she must guess a likelihood \( p^i(a_i^{-i}|\theta^i) \) with parameters \( \theta^i \). While many candidates for \( p^i \) would be thinkable, for this application I specify \( p^i \) as a Bernoulli distribution. This implies one unknown parameter \( \theta^i = p^{-i}(a_i^{-i,1}) \), corresponding to the probability that player \(-i\) plays \( a_i^{-i,1} \). Note that subjective beliefs about probabilities of the opponent’s play are distinct from objective probabilities of the opponent’s play by a missing \( t \)-script, which means that players do not take into account that other players learn over time.\(^{12}\)

Player i has to form beliefs \( \mu^i(p^{-i}(a_i^{-i,1}|.) \) about the unknown parameter \( p^{-i}(a_i^{-i,1}) \). While beliefs can in principle be specified as an arbitrary probability distribution, I assume that \( \mu^i(p^{-i}(a_i^{-i,1}|.) \) corresponds to a beta distribution:

\[
p^{-i}(a_i^{-i,1}) \sim \mathcal{B}(\alpha_{i,t-1}, \beta_{i,t-1})
\]

where \( \alpha_{i,t-1} \) represents the number of (both observed and hypothetical) trials in which the opponent \(-i\) indeed played \( a_i^{-i,1} \). \( \beta_{i,t-1} \) represents the number of (both observed and hypothetical) trials in which the opponent played \( a_i^{-i,2} \). The priors for the parameters of the beta distribution, \( \alpha_0^{-i}, \beta_0^{-i} \), reflect the numbers trials, in which the opponent played \( a_i^{-i,1} \) or \( a_i^{-i,2} \) respectively, in a hypothetical sample that players have in mind before starting to play the actual game. \( \alpha_0^{-i}, \beta_0^{-i} \) contain two pieces of information: the odds ratio that the opponent plays \( a_i^{-i,1} \) as opposed to \( a_i^{-i,2} \), which is reflected by the ratio \( \frac{\alpha_0^{-i}}{\beta_0^{-i}} \) as well as the magnitude of \( \alpha_0^{-i} \) and \( \beta_0^{-i} \). A high magnitude of \( \alpha_0^{-i} \) and \( \beta_0^{-i} \) reflects a high degree of confidence in the priors and the observed play of the opponent in the game plays less of a role. Conversely, if the magnitude of \( \alpha_0^{-i} \) and \( \beta_0^{-i} \) is low, the 1’s that are added during the play carry more weight.

A beta distribution is chosen for several reasons: first, it is the conjugate prior of the Bernoulli distribution, meaning that under a Bernoulli distributed outcome the posterior distribution is of the same family as the prior distribution. Second, the beta distribution is truncated to the unit interval so that it seems a natural choice for the distribution of a probability value.\(^{13}\)

\(^{12}\)Possible reasons may be cognitive costs or overconfidence, meaning that the player assumes that she is more sophisticated than her opponent. (See e.g. Camerer et al. (2004).)
Choosing an action  A rational learner would simply use the mean from the posterior distribution as an estimate for $p^{-i}(a^{-i,1})$ so that

$$E_{t-1}^R p^{-i}(a^{-i,1}) = \frac{\alpha_{t-1}^{-i}}{\alpha_{t-1}^{-i} + \beta_{t-1}^{-i}}$$

(15)

This has been known as fictitious play (Brown, 1951) in the literature and would result in a deterministic choice conditional on the history of data. However, under Thompson sampling agents make a random draw from the posterior. The posterior conditional on the history up to period $t-1$ is given by (14), from which the agent makes a random draw denoted by $\tilde{p}^{-i}_t(a^{-i,1})$. Once agents made this draw, the choice of the player can be determined as being the optimal one conditional on $\tilde{p}^{-i}_t(a^{-i,1})$.

The player uses her estimate of the probability that the other player plays $a^{-i,1}$, $\tilde{p}^{-i}_t(a^{-i,1})$, to determine whether she herself plays $a^{i,1}$ or $a^{i,2}$. Without loss of generality, consider the row player ROW, using the payoffs from table 2: With $\tilde{p}^i_t COL(L)$ as her estimate for the column player to play $L$, the row player’s discounted expected payoffs are, for instance, if she plays $L$:

$$E(u_{t}^{ROW} | a_{t}^{ROW} = L, \tilde{p}^i_t COL(L)) = P_1^{ROW} \cdot \tilde{p}^i_t COL(L) + P_2^{ROW} \cdot (1 - \tilde{p}^i_t COL(L))$$

(16)

Hence, it is easy to see that the row player plays $L$ if $E(u_{t}^{ROW} | a_{t}^{ROW} = L, \tilde{p}^i_t COL(L)) > E(u_{t}^{ROW} | a_{t}^{ROW} = R, \tilde{p}^i_t COL(L))$ and $R$ if $E(u_{t}^{ROW} | a_{t}^{ROW} = R, \tilde{p}^i_t COL(L)) > E(u_{t}^{ROW} | a_{t}^{ROW} = L, \tilde{p}^i_t COL(L))$.

The probability that the row player’s choice $a^i(t)$ is $L$ is given by

$$P^{ROW}_t (L) = \begin{cases} 1 - I_p(a^i_{t-1}, b^i_{t-1}) & \text{if } P_1^{ROW} - P_2^{ROW} - P_3^{ROW} + P_4^{ROW} > 0 \\ I_p(a^i_{t-1}, b^i_{t-1}) & \text{otherwise} \end{cases}$$

(17)

where $I_p$ is the “regularized incomplete beta function”, the c.d.f. of the beta function.

Belief updating  After both players have made their choices, those choices are observed by the other player -i. Hence, every player $i$ uses these observations to update her old beliefs $\mu^i(p^{-i}(a^{-i,1})|\zeta_t^i)$. Belief updating is assumed to be purely Bayesian as in Thompson’s (1933) proposal.

A specific property of the beta distribution is that Bayesian updating implies adding 1 to $\alpha_{t-1}^{-i}$,
if she observes the opponent playing $a_{-i,1}^t$ at the end of period $t-1$ and adding 1 to $\beta_{t-1}^{i}$, if she observes the opponent playing $a_{-i,2}^t$ at the end of period $t-1$. Hence:

$$
\alpha_{t}^{i} = \begin{cases} 
\alpha_{t-1}^{i} + 1 & \text{if } a_{t}^{i} = a_{-i,1}^t \\
\alpha_{t-1}^{i} & \text{if } a_{t}^{i} = a_{-i,2}^t
\end{cases} 
$$  \hspace{1cm} (18)

$$
\beta_{t}^{i} = \begin{cases} 
\beta_{t-1}^{i} + 1 & \text{if } a_{t}^{i} = a_{-i,1}^t \\
\beta_{t-1}^{i} & \text{if } a_{t}^{i} = a_{-i,2}^t
\end{cases} 
$$  \hspace{1cm} (19)

The beta distribution $\mathcal{B}(\alpha_t^{i}, \beta_t^{i})$ constitutes player $i$’s posterior belief $\mu^i(p^{-i}(a_{-i,1}^t)|z_{t+1}^i)$.

### 3.2 Empirical evaluation

#### 3.2.1 Dataset

10 constant-sum games from Erev et al. (2007) are taken (payoffs shown in table 3), in which each subject played 500 periods of a 2x2 game against the same opponent. The numbers in each cell represent the probabilities that players win a fixed lottery prize $v$, set to $0.04$, on each trial. For instance, if ROW plays T and COL plays L, player 1 will win $v$ with the specified probability $P_{1}^{ROW}$, while player 2 will win $v$ with probability $P_{1}^{COL} \equiv 1 - P_{1}^{ROW}$. Such a design has the advantage to control for risk preferences (see e.g. Roth and Malouf (1979).) Each player knew the probabilities in the payoff matrix in the game she played. She was also informed about the action the other player has chosen and therefore of the probability with which her opponent won the lottery prize. The player also knew whether or not she herself received the lottery. However, players were not informed whether the opponent received the lottery prize or not.

The dynamic patterns of four of those games are shown in figure 1. It is of particular interest whether the noise patterns are similar across those ten games. Sign tests do not reject the hypothesis that the variance in the first 75% of the rounds is the same as in the last 25% of the rounds for nine out of ten games. The exception is game 1, displaying weak evidence of declining variance for the row player (p-value: 0.0898) and stronger evidence for the column player (p-value: 0.0195).

Different degrees of volatility across games are expected due to different Nash equilibria. For example, if the Nash equilibrium is 50-50, then equilibrium play, involving play of both strategies
Table 3: Games in Erev et al. (2007)

with equal frequencies, has a greater variance than for a Nash equilibrium 80-20, where one
strategy is played much more often than the other.

3.2.2 Methodology

The initial conditions (or priors) are estimated together with the rationality parameter \( \lambda \). The
initial priors are assumed to be the same within every type of player to save degrees of freedom.
Moreover, the parameters are assumed to be stationary over time, which can be considered rea-
sonable, as the environment to which the subjects are exposed in the dataset is stationary over
time by design of the experimenter (apart from other agents’ behavior, which is endogenous in
the behavioral models I consider.)

One could estimate the parameters using the entire sample. However, numerous studies have
shown that in-sample predictive ability does not necessarily imply out-of-sample predictive ability
(Welch and Goyal, 2008; Stock and Watson, 2003) also because there is the peril of overfitting (see
e.g. Leamer (1978)), meaning that the parameter estimates are driven by noise in the calibration
dataset. Thus, the ultimate test of a model is out-of-sample fitting, which is a common method-
ology both in microeconomics (e.g. Camerer and Ho (1999)) and macroeconomics (see Clark and
McCracken (2013)).

<table>
<thead>
<tr>
<th>Game</th>
<th>Payoff matrix</th>
<th>Game</th>
<th>Payoff matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L  R</td>
<td></td>
<td>L  R</td>
</tr>
<tr>
<td>1</td>
<td>T .77 .35</td>
<td>6</td>
<td>T .46 .54</td>
</tr>
<tr>
<td></td>
<td>B .08 .48</td>
<td></td>
<td>B .61 .23</td>
</tr>
<tr>
<td>2</td>
<td>T .73 .74</td>
<td>7</td>
<td>T .89 .53</td>
</tr>
<tr>
<td></td>
<td>B .87 .20</td>
<td></td>
<td>B .82 .92</td>
</tr>
<tr>
<td>3</td>
<td>T .63 .08</td>
<td>8</td>
<td>T .88 .38</td>
</tr>
<tr>
<td></td>
<td>B .01 .17</td>
<td></td>
<td>B .40 .55</td>
</tr>
<tr>
<td>4</td>
<td>T .55 .75</td>
<td>9</td>
<td>T .40 .76</td>
</tr>
<tr>
<td></td>
<td>B .73 .60</td>
<td></td>
<td>B .91 .23</td>
</tr>
<tr>
<td>5</td>
<td>T .50 .64</td>
<td>10</td>
<td>T .69 .05</td>
</tr>
<tr>
<td></td>
<td>B .93 .40</td>
<td></td>
<td>B .13 .33</td>
</tr>
</tbody>
</table>
To hedge against overfitting, a cross-validation procedure is adopted. This means that the sample is divided into two parts: one part is the training sample, being used to estimate the model parameters. Those estimates are then used to predict the datapoints of the second part of the sample, the validation sample. Cross-validation is a powerful tool, because several partitions can be used and the results of several validation usually make conclusions about model evaluation more robust.

I use the log-likelihood (LL) as a loss function:

$$LL = \sum_{i=1}^{T} \sum_{i=1}^{n_v} \ln(f(a^i_t))$$  \hspace{1cm} (20)$$

where $n_v$ denotes the number of subjects in the validation sample. The likelihood is an appropriate measure, since the models provide density forecasts due to their stochastic nature. Density forecasts give forecasts of all values that the variable of interest can take with a likelihood measure.

There are two empirical approaches to learning: first, there are individual learning models
focusing on individual behavior in the learning process. The likelihood as in (20) already provides a suitable loss function for evaluating a model’s suitability to predict individual learning. The second approach consists of population models, making predictions about the aggregate behavior of a population. I consider the empirical power of Thompson sampling for population models by adopting a weighted likelihood approach (Amisano and Giacomini, 2007) in which each observation is weighed by a weighting function $w(.)$ in the joint likelihood:

$$WL = \sum_{t=1}^{T} \sum_{i=1}^{n_v} w(a_i^t) \ln(f(a_i^t))$$

(21)

The log-likelihood specification in (20) is a special case of (21), applying equal weight $w(a_i^t) = 1$ to each observation. To evaluate the adequacy of Thompson sampling as a population learning model, I choose a weighting function that attaches high weight on observations that are close to the cross-sectional mean in the specific period $t$ and low weight on observation in the tails. More formally:

$$WL = \sum_{t=P+1}^{T} \sum_{i=1}^{N} w\left(\frac{a_i^t - \bar{a}_t}{\sigma}\right) \ln(f(a_i^t))$$

(22)

where $\sigma$ is the standard deviation of $a_i^t - \bar{a}_t$ in the training sample. In specifying the functional form of the weighting function $w(.)$, I follow Amisano and Giacomini (2007) who suggest a standard normal distribution for a weighting function putting emphasis on observations in the center of the distribution.

A further question is whether a generalization criterion (see Busemeyer and Wang (2000)) should be used so that the learning parameters are not only stable over time but also stable over games. If the learning parameters differed a lot across games, the natural question that would arise would be: what drives this difference in the learning parameters? Thus, it would be desirable to obtain estimates that are stable over setups so that one could predict the behavior in games a priori before collecting data. The conjecture that the learning parameters should be stable over environments provides a motivation for using the 10 different games as 10 partitions of the sample so that 10-fold cross-validation is used. This means that the exogenous parameters are estimated nine times, always leaving out one of the games. The game left out is then used for pseudo out-of-sample prediction.
QRE  Applying QRE is straightforward. The logit equations constitute two equations with two unknown probabilities for every $\lambda$. Solving for these probabilities can be included in any search algorithm, finding the $\lambda$-parameter, which maximizes the log-likelihood of the training sample.

TS  Since the payoff matrices are asymmetric, the priors for TS are allowed to differ for ROW and COL. However, following e.g. Camerer and Ho (1999), the priors were restricted to be the same across all players. Thus, Thompson sampling has four free parameters corresponding to the priors $\alpha_{0}^{ROW}, \beta_{0}^{ROW}, \alpha_{0}^{COL}, \beta_{0}^{COL}$. The data display stark differences in initial play, which can plausibly attributed to different information given to the subjects before the start of the game (e.g. payoff matrices). It would thus be incorrect to use the initial conditions of one game to predict the dynamics of another game. Hence, in a first step the initial priors have been estimated independently for every game with the objective to find a theory how initial play is determined. The relative frequencies $\frac{\alpha_{i}^{-1}}{\beta_{0}^{-1}}$ have been found to be close to the Nash equilibria of every game. Hence, the restrictions $\alpha_{0}^{-1} = p_{-i,*} \cdot N$ and $\beta_{0}^{-1} = (1 - p_{-i,*}) \cdot N$ respectively have been imposed, where $p_{-i,*}$ denotes the equilibrium play of L or T depending on the opponent and $N$ the size of the “hypothetical” sample that players have in mind before playing.

Logit  As the logit approach uses the same error structure as QRE (with one exogenous parameter) but the Bernouilli specification of TS (four exogenous parameters), the logit approach has five free parameters: the priors $\alpha_{0}^{ROW}, \beta_{0}^{ROW}, \alpha_{0}^{COL}, \beta_{0}^{COL}$ as in Thompson sampling as well as the $\lambda$-parameter. When independently estimated over games, the initial priors have also been found to be reasonably close to the Nash equilibria. Therefore and for the sake of comparability, the same restrictions as for Thompson sampling have been applied for logit. This gives two exogenous parameters to estimate: $\lambda$ and the “hypothetical” sample size $N$.

3.2.3 Estimation results

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
<th>Nash</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>Tie (0.0588)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>QRE (0.0069)</td>
<td>QRE (0.0093)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>TS (0.0051)</td>
<td>Logit (0.0069)</td>
<td>QRE (0.0051)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nash</td>
<td>Tie (0.2041)</td>
<td>Tie (0.1676)</td>
<td>QRE (0.0124)</td>
<td>Tie (0.4473)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: 2x2 games and individual behavior: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)
Model comparison  Wilcoxon signed-rank tests have been conducted in order to compare the performance of the models pairwise. Table 4 reports the results for individual behavior and table 5 for average behavior.

Observation 1. While all models do better than random decision-making, only QRE does significantly better than Nash equilibrium.

Table 5: 2x2 games and average behavior: Preferred model by the Wilcoxon signed-rank test in pairwise comparison (p-values in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
<th>Nash</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>Tie (0.6455)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>Tie (0.2041)</td>
<td>QRE (0.0093)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>Tie (0.0051)</td>
<td>Logit (0.0164)</td>
<td>QRE (0.0093)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>Tie (0.6455)</td>
<td>Tie (0.5755)</td>
<td>Tie (0.8808)</td>
<td>Nash (0.0164)</td>
<td>-</td>
</tr>
<tr>
<td>Nash</td>
<td>Tie (0.6455)</td>
<td>Tie (0.5755)</td>
<td>Tie (0.8808)</td>
<td>Nash (0.0164)</td>
<td>-</td>
</tr>
</tbody>
</table>

The observations for individual behavior are altogether robust for average behavior as indicated by table 5. The only exceptions are that QRE does not do significantly better than Nash.
equilibrium and TS. Since there are no economically significant differences in the randomness observed in those games, these results are not surprising.

**Parameters** Table 6 reports the results.

**Observation 2.** The priors for successes and failures TS take very high values, while they take low values for the Bayesian logit.

The magnitude of $\alpha_{-i}^0$ and $\beta_{-i}^0$ reflect the sizes of a hypothetical sample before playing the game. A large number implies a high degree of confidence in the priors and the observed play of the opponent in the game plays less of a role. Conversely, if the magnitude of $\alpha_{-i}^0$ and $\beta_{-i}^0$ is low, the 1’s that are added during the play carry more weight.

For TS, $\alpha_{-i}^0$ and $\beta_{-i}^0$ have particularly high magnitudes. This implies that individuals hardly revise their beliefs through the behavior observed from the opponent. These results hence postulate only a limited role of learning and that play is rather determined by (almost) fixed beliefs at the beginning of the respective game. This can be explained by the observation that in this dataset behavior in 2x2 games frequently shows a particular degree of inertia.

For the Bayesian logit, the initial sample takes rather low values, being about 4 on average. The difference to TS can be explained by the fact that the Bayesian logit allows large shocks in the action space, while TS only allows for shocks in the beliefs.

<table>
<thead>
<tr>
<th>Size of hypothetical prior sample</th>
<th>Rationality parameter $\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit $N$</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td>(1.15)</td>
</tr>
<tr>
<td>TS $432,597.88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(40,824.77)</td>
</tr>
<tr>
<td>QRE</td>
<td>6.31</td>
</tr>
</tbody>
</table>

standard errors in parentheses

Table 6: MLE estimates

4 **Application to expectation formation**

Experiments have shown that, even in the presence of a unique equilibrium, learning dynamics on a continuous strategy space depend on the kind of feedback in the underlying system.
and Fehr, 2006; Hommes, 2013): if there is negative feedback, i.e. choices are strategic substitutes, dynamics converge very likely and fast to the equilibrium; if there is positive feedback, i.e. choices are strategic complements, dynamics converge less likely and slowly to the equilibrium. The reason is that under strategic substitutes, agents are induced to choose opposite actions to other agents. Hence, rational agents would do the opposite of less rational agents and thus eventually dominate the market. Conversely, under strategic complements, agents are induced to coordinate, so that rational agents are induced to mimic less rational agents. Hence, non-rational behavior can dominate the market.

A challenge in the literature was to find a model that endogenously predicts these dynamics. Anufriev and Hommes (2012) propose a heterogeneous-agent model where agents endogenously choose between different forecasting rules, and according to Hommes (2013), a homogeneous forecasting rule that endogenously predicts these dynamics is yet to be found. Thompson sampling is a tractable, homogeneous behavioral rule that can fill this gap. Using the experimental data by Heemeijer et al. (2009), I show that Thompson sampling can predict both the dynamics in both kind of situations: situations with negative feedback like the Cobweb model and situations with positive feedback like an asset market.

4.1 Dataset

I take the dataset of Heemeijer et al. (2009). The setup is a learning-to-forecast game in the spirit of Marimon and Sunder (1994), where subjects are only paid for their forecasting performance but market outcomes are determined by the computer. \( n = 6 \) participants are asked to form beliefs about the realization of a market price for 50 periods. After the six participants have typed their beliefs for the price in period \( t \) into the computer interface, the mean over the individual beliefs \( \bar{p}^{e,i}_t \) (corresponding to the actions) for the price realization in period \( t \) are inserted into a price adjustment equation:

\[
p_t = c + b \frac{1}{n} \sum_{i=1}^{n} p^{e,i}_t + \epsilon_t
\]

(23)

c, b represent exogenous parameters and \( \epsilon_t \sim N(0, \frac{1}{4}) \) represents a stochastic shock. Equation (23) (including the shock realizations \( \epsilon_t \)) is never disclosed to the participants. Heemeijer et al. (2009) calibrate the parameters \( c, b \) different in their two treatments: in the treatment with strategic
complements, the parameters are set to $c = \frac{20}{21} \cdot 3$ and $b = \frac{20}{21}$, while the treatment with strategic substitutes, these parameters are set to $c = \frac{20}{21} \cdot 123$ and $b = -\frac{20}{21}$. Heemeijer et al. (2009) show that the pricing equation under strategic substitutes can be derived as the reduced-form equation of a Cobweb model, while the pricing equation under strategic complements can be derived as the reduced-form equation of an asset market setup.\footnote{In fact, an asset market and a cobweb model can both be considered to be a time-overlapping beauty contest game with an interior solution.}

It can easily be verified that this implies a unique fundamental $p^f = 60$ as well as a unique rational expectations equilibrium.

Participants are rewarded according to a quadratic distance equation:

$$u_t^i = \max \{0, 1300 - \frac{1300}{49} (p_t - p_t^{e,i})^2\} \quad (24)$$

The results of the different experimental groups are displayed in figures 4 and 5. For strategic substitutes, Heemeijer et al. (2009) find quick convergence to the fundamental, while for strategic complements, there is no evidence for convergence\footnote{The explosive dynamics in group 5 of the strategic complements treatment are caused by one individual forecasting 5250 for period 8. This likely represent a typo and the subject may have intended to forecast 52.50. Because of this outlier, group 5 has been excluded from the estimation.} However, coordination in the strategic complement treatment occurs fast, while it is slow in the strategic substitutes treatment. Heemeijer et al. (2009) find a significantly higher standard deviation for periods 2-7 in the strategic substitute treatment than in the strategic complement treatment. After period 7, coordination is high in both treatments.

### 4.2 Theory

Suppose agents do not know the price adjustment equation (23). Following Evans and Honkapohja (2001), I assume that they nevertheless perceive the law of motion of the prices to follow the same functional form as under rational expectations. However, consistently with the information structure in the experiment, they do not know the parameters. This means that players perceive that the prices $p_t$ are drawn from a normal distribution with a fixed mean $p^*$ so that the perception for player i of the form:

$$p_t \sim N(p^*, \sigma^2) \quad (25)$$
\( p^* \) is unknown and corresponds to the variable or state \( \theta^{*i} (\forall i \in \mathcal{I}) \) that needs to be learned. For technical simplicity, I assume that the variance \( \sigma^2 \) is known (or the player believes to know the variance). Yet, I allow for the possibility that the variance \( \sigma^2 \) is not necessarily the same as the variance of the exogenous shocks, as players may anticipate the shocks that are introduced by the behavior of other players. Hence, one can think of a shock \( \nu_t \), being different from the \( \epsilon \)-shock, so that \( \nu_t \sim N(0, \sigma^2) \).

Consider a period \( t \) where each player \( i \) needs to forecast \( p_t \) given past data until period \( t-1 \).

Given (25), the optimal forecast is \( \mathbb{E}_t p_t = p^* \). However, the challenge is that \( p^* \) is unknown. The price realizations \( p_t \) are observed by the player and thus constitute “signals” or “measurements” of \( p^* \).

Since \( p^* \) is unknown, it can be considered as stochastic. The player has to form an initial, prior belief about it. In principle, this prior could take any form. However, for technical simplicity, it is assumed to be Gaussian:

\[
p^*_t \sim N(\bar{p}^*_t, \rho_t)
\]

where \( p^*_t \) denotes the candidate value for \( p^* \) at time \( t \). An optimizing Bayesian agent would forecast \( \mathbb{E}_t p_t = \mathbb{E}_t p^*_t = \bar{p}^*_t \). However, an agent using Thompson sampling proceeds through the steps 1-4 in the previous section. To instantiate her belief about \( p^* \), she makes a random draw \( \tilde{p}^*_t \) from (26).

Conditionally on \( \tilde{p}^*_t \), she chooses \( p^{e,i}_t \) as to maximize her expected reward. Using equation (24), the expected reward is given by:

\[
\mathbb{E}_t(u^i_t|\tilde{p}^*_t) = \mathbb{E}_t[\max\{0, 1300 - \frac{1300}{49} (\tilde{p}^*_t + \nu_t - p^{e,i}_t)^2\}]
\]

which implies an optimal action of \( p^{e,i}_t = \tilde{p}^*_t \). Once every individual has made her choice, the average price forecast in period 1, \( p^*_t \equiv \frac{1}{n} \sum_{i=1}^{n} p^{e,i}_t \), can be obtained and \( p_t \) can calculated and announced to the players.

With \( p_t \), the distribution in equation (26) can be updated in a Bayesian way to obtain \( p^*_{t+1} \sim N(\bar{p}^*_{t+1}, \rho_{t+1}) \). The Bayesian update is given by the Kalman filter (Kalman, 1960). The Kalman

\footnote{\( p_t \) is yet unknown at the beginning of period \( t \), as it depends on \( p^{e,i}_t \), the forecast that has to be submitted in period \( t \).}
filter uses a filtering equation of the form:

\[ \bar{p}_{t+1}^* = \bar{p}_t^* + g_{t+1}(p_t - \bar{p}_t^*) \]  

(28)

Bayesian learning optimally determines \( g_t \) using the Kalman filter (Kalman, 1960), which minimizes the expected loss between the price to be forecast, \( p_{t+2} \), and the posterior mean, \( \bar{p}_{t+1}^* \):

\[ \kappa_{t+1} \equiv \arg \min_{g_{t+1}} \mathbb{E}[(\bar{p}_t^* + g_{t+1}(p_t - \bar{p}_t^*)) - p_{t+1}]^2 = \frac{\rho_t}{\rho_t + \sigma^2} \]  

(29)

The variance of \( \bar{p}_{t+1}^* \), denoted by \( \rho_{t+1} \), is obtained as

\[ \rho_{t+1} = \left( \frac{1}{\rho_t} + \frac{1}{\sigma^2} \right)^{-1} \]  

(30)

To initiate the individual belief for period \( t + 1 \), the distribution \( p_{t+1}^* \sim N(\bar{p}_{t+1}, \rho_{t+1}) \) is used and the same steps are repeated in further periods.

To uniquely determine the distributions that generate the individual beliefs, the prior mean \( \bar{p}_0^* \), the prior variance \( \rho_0 \) and the variance of the perceived shock, \( \sigma^2 \), need to be calibrated.

![Figure 4: Strategic substitutability sessions from Heemeijer et al. (2009)](image1)

![Figure 5: Strategic complementary sessions from Heemeijer et al. (2009)](image2)

### 4.3 Methodology

The methodology to empirically evaluate Thompson sampling relative to other approaches of endogenous noise is similar to the one used for 2x2 games. A cross-validation strategy is employed, in which the data is divided into \( k \) independent subsamples, of which \( k-1 \) are used as a training sample to estimate the parameters and one subsample is used for validation. I use every indepen-
dent experimental group as one subsample. Having identified one group as an outlier that would distort the estimation, this gives \( k = 12 \) subsamples. Since a Kolmogorov-Smirnov test does not reject the hypothesis that first-period play follows the same distribution (p-value: 0.538) for both treatments, one can plausibly assume that initial play is independent of the treatment.

To compute the likelihood of QRE and Logit on the continuous strategy space numerically, a discrete approximation to the continuous functions has been provided, i.e. the space has been divided up into bins so that each bin can be mapped to an action and thus a probability of this bin being chosen. I divide the space up into 100 equal intervals \([0, 1), [1, 2), \ldots, [99, 100]\). The specifics of how each model is applied in every case are exemplified in the following subsections.

**QRE**

To the best of my knowledge, this is the first application of QRE to a learning-to-forecast design. A preliminary question is whether \( \lambda \) is stable over time. In a multinomial logit regression on time dummies, using a categorical variable containing each unit from 0-100 as the dependent variable, this hypothesis could not be rejected. To make the calculation simpler, I assume the midpoint of each interval is used for payoff consideration, which is \( p^e_{ij} - 0.5 \) for \( p^e_{ij} = 1, 2, 3, \ldots, 100 \).

With the payoff function given in (2), the expected payoff is

\[
E(u_t | p^e_{ij} = p^c_{ij}, Pr(p^e_t)) = \\
\sum_{p^{c} = 1}^{100} Pr(p^e_t = p^c) \cdot \left(1 - \frac{1}{49} \left((c + b \cdot (p^c - 0.5) - (p^e_{ij} - 0.5))^2 + 0.25\right)\right)
\]  

(31)

where the last term comes from the exogenous disturbance \( \epsilon_t \sim N(0, 0.25) \). For every \( \lambda \), a fixed point is defined for the probability distribution \( Pr(p^e_{ij} = p^c_{ij}) \). This fixed point is obtained through value function iteration. Since (31) contains the average action \( p^c \) and the convolution (distribution of the sum of random variables) of a logit rule does not have a closed-form solution, the probability distribution of the average was simulated in every iteration of the search algorithm.

---

16 This requires caution, since a discretized space implies a probability mass function for QRE and Logit instead of a density function. However, if the bins are chosen to be of equal size, the probability of each strategy corresponds to its density so that \( Pr(a_t) = f(a_t) \). Since under the division of bins used here, QRE, Logit and TS all create densities, their log-likelihoods are comparable.

17 Less than 1% of all price forecasts are greater than 100.

18 QRE has previously been applied to the p-beauty contest game by [Breitmoser (2012)](Breitmoser2012), who uses a similar methodology to this paper.

19 Following other examples in the experimental literature such as [Anufriev and Hommes (2012)](Anufriev2012), I assume that subjects consider the payoffs without the truncation at zero for the sake of analytical tractability.

20 For the treatments with the robot traders, \( p^c \) takes into account the choices of these computerized traders. Taking into account \( n_t \) explicitly would require solving the fixed point problem for a wide range of values for \( p_{t-1} \) and thus makes the computational problem disproportionately more burdensome.
by making 6 draws from the estimate of the probability distribution in the current iteration of
the algorithm, calculating the mean and repeating that procedure 2,000 times. The obtained
frequencies for the mean can then be used as a good approximation of the distribution of the
mean choice. The approximated distribution was then inserted into the right-hand side of the
logit rule to calculate the individual probabilities. The algorithm would stop once this resulting
probability is consistent with the probability used for simulating the distribution. The \( \lambda \) that
maximizes the log-likelihood was yielded by embedding the fixed-point problem into a derivative-
free simplex search algorithm. (Nelder and Mead [1965])

**TS** As exemplified in section [4], there are three prior parameters that need to be calibrated for
the purpose of Bayesian updating: \( \bar{p}_0, \rho_0, \sigma^2 \).

**Logit** Agents have the same Gaussian perception as in Thompson sampling:

\[
p_t \sim N(p^*, \sigma^2)
\]

(32)

with \( \nu_t \sim N(0, \sigma^2) \). This perception (or belief structure) is used for the calculation of the expected
payoff for a particular forecast \( p^{e,i} \), \( V(.) = \mathbb{E}(u_t | p_t = p^* + \nu_t, p_t^{c,i} = p^{e,i}) \), in the logit expression,
where the payoff is given by equation (24).

**Proposition 3.** The expected payoff for a particular forecast \( p^{e,i} \) conditional on the random walk
perception in (32) is given by

\[
\mathbb{E}_t(u_t | p_t = p^* + \nu_t, p_t^{c,i} = p^{e,i}) = 1300 - \frac{1300}{49} [\sigma^2 + \rho_t + \bar{p}^2_t - 2\bar{p}^*tp^{e,i} + p^{e,i2}]
\]

(33)

where \( \bar{p}^*_t \) denotes the expectation of \( p^* \) and \( \rho_t \) the posterior variance using information up to period
t-1 [21]

**Proof.** See Appendix [6, 1] \( \square \)

This expected payoff is inserted into the logit equation. The integral of \( \exp(\lambda \mathbb{E}_t(u_t | p_t = p^* + \nu_t, p_t^{c,i} = p^{e,i})) \) has been evaluated by a discrete approximation, dividing the strategy space
from 0 to 100 up into equal bins [22]. Altogether, three parameters need to be estimated for the

[21] The max(.) is ignored here for analytical tractability.
[22] The midpoint of every interval has been used for expected payoff calculation.
Bayesian logit: the two initial priors from Thompson sampling, $\bar{p}_0^*, \rho_0$, the perceived noise variance $\sigma^2$, as well as the rationality parameter $\lambda$ from the logit distribution.

### 4.4 Estimation results

**Model comparison**  
Rational expectations has not been used as a benchmark, since it implies deterministic choices, which would correspond to a likelihood of zero for the slightest deviation. Figures 6 and 7 show the results for individual and average behavior respectively and tables 7 and 8 show the result for the pairwise Wilcoxon signed-rank tests. Altogether, the results are very similar for individual and average behavior. The following observations stand out:

**Observation 4.** All three models perform significantly better than a random draw on the interval $[0,100]$.

**Observation 5.** The empirical fit of Thompson sampling and the Bayesian logit is very similar.

**Observation 6.** Both Thompson sampling and the Bayesian logit provide a better fit than QRE.

Although the empirical fit of Thompson sampling and the Bayesian logit is very similar, the fact that Thompson sampling is slightly better in 11 out of 12 groups renders it the preferred model by the signed-rank test.

Note that $\rho_t, \sigma^2$ do not only appear as constants in the expected payoff but also in the Kalman gain so that they determine $\bar{p}_t^*$.

---

23 Note that $\rho_t, \sigma^2$ do not only appear as constants in the expected payoff but also in the Kalman gain so that they determine $\bar{p}_t^*$. 

28
Both the Bayesian logit and Thompson sampling predict significantly better than QRE. Since the characteristic feature of QRE is equilibrium beliefs, this finding can be interpreted as evidence against equilibrium beliefs.

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>TS (0.0121)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>TS (0.0121)</td>
<td>Logit (0.0278)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>TS (0.0029)</td>
<td>Logit (0.0048)</td>
<td>QRE (0.0029)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7: Learning to forecast and individual behavior: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>TS (0.0022)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>TS (0.0076)</td>
<td>Logit (0.0278)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>TS (0.0029)</td>
<td>Logit (0.0029)</td>
<td>QRE (0.0029)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8: Learning to forecast and average behavior: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

**Parameters**

**Observation 7.** For TS, the initial prior mean $\bar{p}_0^*$ is near the fundamental value.

The prior mean $\bar{p}_0^*$ in the logit specification is closer to the initial prices of most groups. Since the value of $\lambda$ implies a high error variance, large shocks are likely to occur so that individual forecasts being very different from the starting prices can easily be explained.

<table>
<thead>
<tr>
<th></th>
<th>prior mean $\bar{\theta}_0$</th>
<th>prior variance $\rho_0$</th>
<th>noise variance $\sigma^2$</th>
<th>rationality $\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>43.03</td>
<td>148.23</td>
<td>26.26</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>(0.88)</td>
<td>(0.47)</td>
<td>(3.64)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>TS</td>
<td>62.38</td>
<td>145.12</td>
<td>1,171.14</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.39)</td>
<td>(10.29)</td>
<td>(133.99)</td>
<td></td>
</tr>
<tr>
<td>QRE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.06)</td>
</tr>
</tbody>
</table>

Table 9: Parameters: Learning-to-forecast experiment
4.5 Endogenous noise variance

As shown by equation 30, the series of posterior variances is deterministically pinned down. As Thompson sampling introduces randomness through sampling from the posterior, there is by design no differences in randomness across environments. Predictions of the dynamics across environments are, however, particularly useful, since they may allow an ex-ante assessment of policies before they are implemented. For example, if a policy can be assessed as undesirable ex-ante, one could both save potential implementation costs and prevent its negative effects. I show that Thompson sampling can be used as a model to make predictions across environments. For the current application, this requires relaxing the assumption that agents know the noise variance $\sigma^2$.

I assume that agents update the noise variance as new observations become available, using the most recent estimate of the mean:

$$
\sigma^2_t \equiv \frac{\psi \sigma^2_0 + \sum_{q=1}^{t-1} (p_q - \bar{p}_{t-1})^2}{\psi + t - 1}
$$

(34)

where $\sigma^2_0$ is the initial prior estimate of the variance at period $t = 0$, based on a hypothetical sample of $\psi$ observations. This corresponds to a Bayesian process of updating the variance, where the mean is treated as known and the conjugate prior for the variance is a scaled inverse $\chi^2$-distribution. (See e.g. Gelman et al. (2013).) This hypothesis of learning about the variance still allows using the Kalman filter.

Hence, this simple modification allows modeling different patterns of across environments in two aspects: first, since equation 30 shows that the posterior variances positively depend on the noise variance, using an estimate that depends on time, $\sigma^2_t$, ensures that randomness can differ across environments; second, equation 29 shows that the Kalman gain negatively depends on the noise variance so that the adjustment speed can also differ across environments.

**Methodology**  Continued use of the Kalman filter allows proceeding in the same way as above as in section 4.3. The main difference is that four parameters need to be estimated for TS ($\bar{p}_0, \rho_0, \psi, \sigma^2_0$) and five for Logit ($\bar{p}_0, \rho_0, \lambda, \psi, \sigma^2_0$). Since the unrestricted estimation yields implausibly high values for the prior variances, the variances have been restricted to 2,500, being the maximum possible variance on the interval $[0, 100]$ (corresponding to the case, in which half of the
observations are 0 and half of the observations are 100.)

4.5.1 Estimation results

Figure 8: Learning to forecast with endogenous variance: Likelihoods of validation samples (individual behavior)

Figure 9: Learning to forecast with endogenous variance: Weighted likelihoods of validation samples (average behavior)

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>Tie (0.1164)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>TS (0.0189)</td>
<td>Logit (0.0048)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>TS (0.0022)</td>
<td>Logit (0.0022)</td>
<td>QRE (0.0029)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 10: Learning to forecast with endogenous variance and individual behavior: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>Logit</th>
<th>QRE</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit</td>
<td>TS (0.0121)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QRE</td>
<td>TS (0.0022)</td>
<td>Logit (0.0029)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>TS (0.0022)</td>
<td>Logit (0.0029)</td>
<td>QRE (0.0029)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 11: Learning to forecast with endogenous variance and average behavior: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

Model comparison Figure 8 and 9 show the likelihood estimates for individual and average behavior respectively, while tables 10 and 11 show the results of pairwise Wilcoxon signed-rank tests.
Observation 8. The empirical fit of TS for individual behavior is at least as good as or better than the Bayesian logit for most groups, but substantially worse in group 3 in the strategic substitute treatment.

While overall the difference in empirical fit between TS and the Bayesian logit is insignificant for strategic substitutes, TS predicts strikingly worse for group 3. This is however driven by only few datapoints and can be explained by outliers. Figure 4 reveals that large shocks from otherwise stable dynamics particularly occur in group 3.

Observation 9. For average behavior, the empirical fit of TS is better than the ones of the two benchmarks.

As shown in figure 9, if less weight is attached to outliers, Thompson sampling performs significantly better than the Bayesian logit. A Wilcoxon signed-rank test over the experimental groups reveals that for average behavior the empirical fit of TS with endogenous noise variance is better than the standard case with constant noise variance. (p-value: 0.0029)

<table>
<thead>
<tr>
<th></th>
<th>prior mean</th>
<th>prior variance</th>
<th>rationality</th>
<th>prior sample size</th>
<th>prior noise variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{\theta}_0$</td>
<td>$\rho_0$</td>
<td>$\lambda$</td>
<td>$\psi$</td>
<td>$\sigma^2_0$</td>
</tr>
<tr>
<td>Logit</td>
<td>45.97</td>
<td>2,500</td>
<td>0.61</td>
<td>2.00</td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(0.00)</td>
<td>(0.06)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>TS</td>
<td>62.33</td>
<td>193.35</td>
<td>-</td>
<td>4.33</td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td>(1.23)</td>
<td>(57.84)</td>
<td></td>
<td>(0.75)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>QRE</td>
<td>-</td>
<td>-</td>
<td>0.27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Table 12: Parameters: Learning-to-forecast with endogenous variance

Parameters As shown in table 12 the estimates for the rationality parameter $\lambda$ are similar to the standard Kalman filtering case. Likewise, the value for the prior mean for the logit approach is robust.

Observation 10. The prior variances for both the logit approach and Thompson sampling are higher than in the standard case.

In the logit approach, the noise variance affects mainly the adjustment speed through the Kalman gain, since the spread of the actions is given by the logit distribution. The fact that both $\rho_0$ and $\sigma^2_0$ attain their maximum possible values is a sign that experience acquired during the game
plays less of a role, since high values of the prior variances mitigate the effect of the experience obtained during the game.

For Thompson sampling, the prior noise variance $\sigma_0^2$ also attains its maximum. Since for Thompson sampling, $\sigma_0^2$ has a direct effect on dispersion in the action space, this can easily be explained by the fact that the variance of individual forecasts remains larger than the variance of the observed outcomes $p_t$. However, the small prior sample size $\psi = 4$ allows for learning about the variance during the game.

5 Discussion

This paper has introduced Thompson sampling, a mechanism that has previously mainly been applied to the bandit problem, as a tractable theory of endogenous randomness into interactive games in economics. By applying Thompson sampling to 2x2 games and learning-to-forecast experiments, it has been shown that Thompson sampling is applicable for very different types of setups. Another virtue of Thompson sampling is the simplicity to implement it for predictive purposes in the context of very different setups. Moreover, a potential advantage of Thompson sampling is that it can produce individual differences without specifying many exogenous parameters.

The empirical result of this paper is that Thompson sampling can explain the emergence of different dynamic patterns significantly better than other models. The fit of Thompson sampling to experimental settings with changing dynamics and noise patterns opens up several directions of future research.

First, this paper only considers Thompson sampling as a positive theory. Hence, future research could investigate whether Thompson has any normative appeal in games.

Second, experimental data have their limitations, as they represent artificial environments with a relatively small sample size. Hence, it would be intriguing to test Thompson sampling for different datasets that may contain observational data.

Third, while agents’ priors have been estimated in this paper and the application to 2x2 games has made use of the equilibria for initialization, Thompson sampling can be combined with other specific theories about how agents form their prior beliefs.

Fourth, Thompson sampling can potentially have numerous policy implications. The fact that
Thompson sampling provides an empirically valid and tractable description of individual beliefs. This may open up several directions for future research: in finance and macroeconomics, policy analysis can be conducted under the assumption that agents’ belief formation process corresponds to Thompson sampling instead of rational expectations. Into the bargain, the implications of Thompson sampling for games, firm and consumer behavior, political economy as well as mechanism design can be explored.

References


6 Appendix

6.1 Proof of Proposition 3

Proof. Ignoring the max(.), the score of the learning to forecast experiment for period t, using (24), is:

\[ u^i_t = 1300 - \frac{1300}{49} (p_t - p_{t \epsilon, i}^e)^2 \]  \hspace{1cm} (35)

Under Gaussian perception given by (32), the score can be written as

\[ u^i_t = 1300 - \frac{1300}{49} (p^* + \nu_t - p_{t \epsilon, i}^e)^2 \]  \hspace{1cm} (36)

The expected value under this perception is

\[ \mathbb{E}_t(u^i_t | p_t = p^* + \nu_t, p_{t \epsilon, i}^e = p_{t \epsilon, i}^e) = 1300 - \frac{1300}{49} \mathbb{E}_t\{(p^* + \nu_t - p_{t \epsilon, i}^e)^2 | p_{t \epsilon, i}^e = p_{t \epsilon, i}^e\} \]  \hspace{1cm} (37)

Focusing only on the term \( \xi_{t+1} \), we obtain

\[ \xi_{t+1} \equiv \mathbb{E}_t\{(p^* + \nu_t - p_{t \epsilon, i}^e)^2 | p_{t \epsilon, i}^e = p_{t \epsilon, i}^e\} \]
\[ = \mathbb{E}_t\{(p^* - p_{t \epsilon, i}^e)^2\} + \mathbb{E}_t\{\nu_t^2\} \]
\[ = \mathbb{E}_t\{p^*^2 - 2p^* \cdot p_{t \epsilon, i}^e + p_{t \epsilon, i}^e^2\} + \sigma^2 \]
\[ = \rho_t + \bar{p}^2_t - 2\bar{p}^* \cdot p_{t \epsilon, i}^e + p_{t \epsilon, i}^e^2 + \sigma^2 \]  \hspace{1cm} (38)

where the last line uses the fact that \( \text{Var}(\mu) = E(\mu^2) - (E(\mu))^2 \). \( \square \)