Mauersberger, Felix

Conference Paper

Thompson Sampling: Endogenously Random Behavior in Games and Markets

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Experimental Economics II, No. B05-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Mauersberger, Felix (2019) : Thompson Sampling: Endogenously Random Behavior in Games and Markets, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Experimental Economics II, No. B05-V1, ZBW - Leibniz-Informationszentrum Wirtschaft, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/203600

Thompson Sampling: Endogenously Random Behavior in Games and Markets

Felix Mauersberger*

University of Bonn

This draft: February 28, 2019

Abstract

Economists tend to assume that agents maximize their expected utility. However, many different experiments have questioned expected utility maximization by showing that human behavior can be characterized as random. This paper proposes Thompson Sampling as a theory of human behavior across very different situations of dynamic strategic interaction in economics. Thompson Sampling means that agents, having limited information about their environments, update their subjective belief distributions in a Bayesian way and subsequently make a random draw from the posterior. Conditional on that random draw, agents optimize. While Bayesian reasoning has often been shown to be at odds with agents' behavior even in simple environments, using data on experimental games, this paper shows that Bayesian sampling as in Thompson's proposal is a better description of agents' decision-making than commonly used theories of decision-making in economics such as Nash equilibrium, standard Bayesian learning and quantal response equilibrium (QRE) - above all in complex environments with many possible actions.

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

*Institute of Applied Microeconomics, University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany, Email: fmauersberger@uni-bonn.de. I would like to thank Larbi Alaoui, José Apesteguia, Jasmina Arifovic, Jess Benhabib, Gabriele Camera, Mark Dean, John Duffy, Evan Friedman, Xavier Gabaix, Robin Hogarth, David Laibson, Gael Le Mens, Rosemarie Nagel, Antonio Penta, Luba Petersen, Hans-Martin von Gaudecker, Thomas Woiczyk, Michael Woodford as well as the participants of the Stony Brook Workshop on Theoretical and Experimental Macroeconomics, the Columbia University Cognition and Decision Laboratory Seminar, the UPF Internal Microeconomics Seminar, the UPF Behavioral Lunch and the CREi Macroeconomic Breakfast for their feedback. Many thanks also to Ido Erev, Peter Heemeijer, Cars Hommes, Jan Tuinstra and Joep Sonnemans for the data.

stochastic choice

JEL Classifications: C91, C92, D84, E37

1 Introduction

Economists tend to believe that agents act optimally, in the sense that they choose the actions that maximizes their expected payoff. A natural question is how agents may acquire such information. Even undergraduate textbooks teach that equilibrium is reached by a long dynamic adjustment process in which players learn about their environment and other agents' strategies that is often described as Bayesian (see e.g. Osborne (2003), p. 132). Different subdisciplines in economics - such as behavioral game theory, experimental economics, behavioral economics - as well as a large literature in psychology attach a lot of attention to this learning process and whether optimal outcomes are attained at least in the long run. There is a myriad of different models where learners are confronted with uncertainty (e.g. McKelvey and Palfrey (1995); Gilboa and Schmeidler (1995); Fudenberg and Levine (1998); Brock and Hommes (1997); Roth and Erev (1998); Camerer and Ho (1999); Anufriev and Hommes (2012)).

Economists usually assume that, under uncertainty, people make decisions by maximizing their expected utility. However, the theory of expected utility maximization did not fare well. Not only have the axioms of expected utility been seriously questioned, but also simple individual decision-making experiments in psychology where the principles of learning were tested reject the notion of expected utility maximization. Many experiments have shown that human behavior can more accurately be characterized as random. These experiments represented simple tasks like predicting whether the next card would be red or blue or whether a light at the end of a tunnel would appear on the left or the right. (Vulkan (2000) for a survey.) The data have repeatedly demonstrated that subjects match the underlying probabilities, i.e. if p denotes the probability that the outcome is left, then subjects would, after an initial learning phase choose Left with probability p. This kind of behavior has not only been found in humans but also in animals like rats, pigeons and bumblebees: instead of choosing source exclusively, animals are found to adopt an adaptive sampling strategy, even though there are opportunity costs to sampling like travel time. (See e.g. Keasar et al. (2002))

The idea of probability matching is old and dates back to Thompson (1933), who proposed it

as a solution to the exploration-exploitation trade-off in bandit-type tasks. Thompson Sampling means that decision-makers randomize actions based on the probability that this action is believed to be optimal when faced with an uncertain environment. Thompson Sampling consists of three steps: first, after obtaining new information, the agents updates her prior subjective probability distribution in a Bayesian way to yield a posterior; second, in lieu of making optimal use of the posterior, which means calculating the expected value, the agent makes a random draw from the posterior; third, the agent treats the random draw as the truth and responds optimally to it.

Bayesian explanations have been popular in cognitive science, since many complex concepts such as perception, (e.g. (Frisby and Stone, 2013)) psychophysics, (e.g. Wolpert (2007)) language, (Chater and Manning, 2006) motor control (Treutwein, 1995) etc. have been successfully modeled in a Bayesian way. On the other hand, many experimental studies have shown that human decisions are at odds with the Bayesian approach; and a large psychological literature shows that the Bayesian approach of representing all possible probabilities and making exact calculations is infeasible for any physical system, including the human brain.[1]. Thompson Sampling resolves this apparent paradox: even if the brain is Bayesian, it need not calculate probabilities. Instead the brain can be interpreted as sampling from the Bayesian posterior. For a Thompson Sampler, knowledge of the entire distribution is not necessary, since it can work with a partition of this posterior distribution. This sampling is consistent with the availability heuristic, i.e. estimating the probability of an event by generating plausible examples in one's mind. (Tversky and Kahneman, 1973) This rule of thumb can be interpreted as a product of Thompson sampling.

This paper considers the use of Thompson Sampling as a descriptive theory for decision-making in economics. The particular appeal of Thompson Sampling is quite multifarious, encompassing both positive and normative as well as computational aspects. Computationally, it is appealing because of its simplicity and tractability, which is why it is nowadays frequently used for online learning problems. On the positive side, it has been shown to describe subjects' behavior well in experimental bandit tasks (Speekenbrink and Konstantinidis, 2015; Gershman, 2018) and other individual decision-making setups (e.g. Wozny et al. (2010).) Furthermore, it is consistent with neuroscientific evidence, because it has been shown that different areas of the brain are activated for learning than for information acquisition, so that it is possible that they are entirely different, e.g. one being optimal, the other one being based on heuristics. (O'Doherty et al., 2004; Behrens

[1]West and Stanovich (2003) provide a evidence for a positive association between cognitive abilities and utility maximization.

et al., 2007; Payzan-LeNestour and Bossaerts, 2015) Importantly, however, Thompson Sampling has normative appeal. While Thompson Sampling is not optimal in the sense that it maximizes expected utility, it has been shown to be regret minimizing in the bandit literature. (May et al., 2012; Agrawal and Goyal, 2012a,b; Kaufmann et al., 2012; Scott, 2010; Chapelle and Li, 2011). In contrast to that, full Bayesian approaches like the Gittins index (Gittins, 1979) display a positive probability of converging to a suboptimal slot machine even when allowing for an exploration phase.[2] (Brezzi and Lai, 2000) Moreover, it has been shown to ensure that the agents asymptotically learns to act optimally in a more general class of stochastic environments that may be non-Markovian, non-ergodic and partially observable. (Leike et al., 2018)

In order to evaluate Thompson Sampling empirically, one needs to consider its falsifiable predictions. Thompson Sampling has the following implications: (1) People's behavior is stochastic due to the random nature of the samples drawn. (2) While there is noise in individual behavior, the average decisions may be correct. This generates the "wisdom of crowds". (3) The distribution of actions is non-stationary over time and across setup, since actions are guided by samples from the posterior. (4) Thompson Sampling should explain behavior in complex tasks particularly well.

(1) has been corroborated by a wide range of experiments and has triggered the development of a large stochastic choice literature, starting with Block and Marshak (1960) and McFadden (1974). Experimental evidence for (2), the "wisdom of crowds"-effect has been provided for example by Nagel and Vriend (1999) and DellaVigna (2018). Since (1) and (2) can be well captured by idiosyncratic exogenous shocks, which have successfully been introduced into a wide range of theories in macroeconomics, microeconomics and econometrics, they would by themselves not warrant the introduction of a new modeling approach like Thompson Sampling. (3) and (4) are far less obvious. Especially (4) can have implications for economics, since markets and the economy in general are large, complex systems of interacting agents.

This paper thus considers datasets where agents strategically interact. Since Thompson Sampling is a cognitive model, I use laboratory data where the cognitively driven randomness from subjects is observed and potential other drivers of randomness are eliminated. Experimental data

[2]To see this intuitively, suppose that the gambler initially assigns priors so that the estimated winning probability of every slot machine is 50 %. Further, suppose that the gambler happens to start playing with a slot machine whose true winning probability is 60 %. Bayesian learning typically ensures that the gambler's estimate of this slot machine's winning probability converges to the true value. As the estimated winning probability of this slot machine would then be higher than the estimated winning probability for any other slot machine, the gambler would keep playing with that slot machine. By doing so, she might neglect the fact that the true winning probability of another slot machine is higher.

on games is consistent with the four hypotheses: first, the randomness observed in games, in particular, has motivated the development of stochastic equilibrium concepts such as quantal response equilibrium (QRE) (McKelvey and Palfrey, 1995; Goeree et al., 2016). Second, the "wisdom of crowds" effect has been observed in games (e.g. (Nagel and Vriend, 1999)). Importantly, data on experimental games also displays evidence for (3), since many experiments have found that the distribution of decisions of human subjects is not stationary over time and across setups (see e.g. (McKelvey and Palfrey, 1995; Dufwenberg et al., 2007)) being reflected by structural breaks in the exogenous parameters of QRE. Data in this literature is also consistent with (4), since subjects clearly learn over time in some setup while they do not learn in other datasets.

Thompson Sampling is tested by evaluating its in-sample fit and predictive ability in two exemplary games that represent varying degrees of complexity: 2x2 games, which can be thought of as "small world", and beauty contest games with incomplete information, which can be thought of as "large world." For the "small world" 2x2 game, I take the large dataset with 10 different constant sum games by (Erev et al., 2002) with full information. For the "large world", I take the dataset by Heemeijer et al. (2009). In this beauty contest experiments with incomplete information, being often called "learning-to-forecast experiment" (Marimon and Sunder, 1994), 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. Subjects are never informed about the functional form of the underlying laws of the market.

In the dataset by Erev et al. (2002), there is a lot of randomness. However, decisions do not converge to the mixed strategy equilibrium over time the patterns of decisions change markedly over time. This can be considered to be evidence of (4). Thompson Sampling is at least found to provide a better fit than the mixed strategy equilibrium and Bayesian learning with shocks. However, Thompson Sampling is not significantly better than QRE.[3] This is unsurprising, since, if agents have difficulty into make the right decisions an environment, their behavior can be interpreted as random and thus be represented by exogenous shocks.

[3]Learning rules such as reinforcement learning (Roth and Erev, 1998) and experience-weighted attraction (Camerer and Ho, 1999) require a discrete action space. While these models can still be applied to continuous decision problems by discretization of the space, the bin size is an arbitrary decision and introduces a dilemma. Small bin sizes have two disadvantages: first, computations are slowed down and have been found not to converge. Second, because only a small part of the action space is "reinforced" after every feedback, even choices that are close to previous choices but do not coincide with them may be penalized by model selection criteria. On the other hand, a big bin size implies an imprecise approximation and imprecise predictions. See Arifovic and Ledyard (2004) for more details. Moreover, Hopkins (2002) shows that reinforcement and fictitious play, being a particular Bayesian model, "are far more similar than were thought."

However, Thompson Sampling delivers both a better in-sample fit and superior predictive ability to QRE and Bayesian learning in the "large world", being represented by the dataset by Heemeijer et al. (2009). In this beauty contest experiments with incomplete information, being often called "learning-to-forecast experiment" (Marimon and Sunder, 1994), 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. Subjects are never informed about the functional form of the underlying laws of the market. 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. In the learning-to-forecast experiment, noise patterns exhibit significant differences both between treatments and over time. Thus the better in-sample fit and the better out-of-sample forecasts of Thompson Sampling are unsurprising.

For a Thompson Sampler knowledge of the entire distribution is not necessary, since it can work with a partition of this posterior distribution.

Thompson Sampling can also explain a number of biases that are well-known in behavioral economics.[4] First of all, it is intuitive that Thompson Sampling can explain biases that have made their way into the vernacular like "cherry picking" (or "confirmation bias") and relying on "anecdotal evidence." Yet, Thompson Sampling can also explain the base-rate fallacy, i.e. the finding that people tend to ignore general information in favor of specifics. Fully taking into account base rates would require exploring the entire probability space, being computationally burdensome. Another well-known bias is the conjunction fallacy, i.e. people usually attach a higher probability to "Linda is a bank teller and is active in the feminist movement" than to "Linda is a bank teller." (Tversky and Kahneman, 1983) This is inconsistent with probability theory, but not inconsistent with Thompson Sampling, since through making details salient the sampler can be guided away from probability peaks. Moreover, Thompson Sampling provides an explanation for the St. Petersburg paradox, i.e. the puzzle that individuals are not willing to pay an infinite amount for a game in which a fair coin is tossed at each stage and the reward is doubled every time heads appears. (Bernoulli, 1954) If individuals sample in their heads, they may draw a small value as an estimate for the winning amount.

Experiments where subjects are asked to predict between left or right or the kind of questions

[4]See Sanborn and Chater (2016) for a detailed survey.

testing biases such as the base-rate neglect and conjunction fallacy represent very simple and intuitive tasks. Yet, violations of basic probability theory are consistently observed. If Bayesian statistics do not work in such simple setups, how will they have a chance to work in richer, more complex setups? Savage (1954) thus proposed restricting Bayesian statistics to "small worlds."

However, Thompson Sampling invalidates this argument, since rich, complex environments are the ones where Thompson Sampling is most effective. Simpler but more abstract problems hinder the search through sampling due to a restriction of the outcome space. A smaller outcome space is associated with less contextual cues, giving fewer hints where to search. Consider the following example (Sanborn and Chater, 2016): it can be extremely challenging to solve a jigsaw puzzle which is uniformly white, while solving a big jigsaw with a colored photograph can be easier, because the sampler can be guided by the context provided and even by past experience from real-world scenarios.

An important implication of Thompson Sampling is that behavior is stochastic due to the random nature of the samples drawn. This can explain the noise in individual behavior, although the average decisions may be correct resulting in the "wisdom of crowds", which has been observed in some experiments. This contrasts to economic models, which are usually deterministic – both equilibrium and non-equilibrium models. However, there is overwhelming evidence in both individual decision-making experiments and in the experimental game theory literature. Randomness in economics has motivated the introduction of random shocks, which are widely used across the field in macroeconomics, econometrics as well as in microeconomics through the use of the random utility model . However, many experiments have found that the distribution of decisions of human subjects is not stationary over time and across setups (see e.g. (McKelvey and Palfrey, 1995; Dufwenberg et al., 2007)) which is why structural breaks have been found in the exogenous parameters of quantal response. An endogenous distribution of decisions is implied by Thompson Sampling, since the probability distribution from which agents sample is updated in a Bayesian way.

This paper thus shows the use of Thompson Sampling as a novel descriptive and predictive theory in economics. The cornerstone of economics can be considered to be strategic interaction with other individuals in markets. This paper thus investigates the empirical fit and the predictive ability of Thompson Sampling to games, which has, to the best of my knowledge, not yet been done. This paper shows that Thompson Sampling is widely applicable to very different setups.

Based on the hypothesis above that Thompson Sampling should be particularly effective in large, complex tasks, Thompson Sampling is applied to two datasets that can be considered extreme opposites: 2x2 games, which can be thought of as "small world", and beauty contest games with incomplete information, which can be thought of as "large world."

For the "small world" 2x2 game, I take the large dataset with 10 different constant sum games by (Erev et al., 2002). There is a lot of randomness, which is inconsistent with the mixed strategy equilibrium. Furthermore, there is no evidence that agents the patterns of decisions change markedly over time. This is unsurprising, as a space with two actions is a small action space, which provides limited scope for changes in the variance of decisions, so that behavior can be well fit with exogenous shocks. Nevertheless, Thompson Sampling is at least found to provide a better fit than the mixed strategy equilibrium and Bayesian learning with shocks. However, Thompson Sampling is not significantly better than QRE.[5]

2 General model

I take the version of Thompson Sampling outlined by Chapelle and Li (2011) and modify it for interactive games. We start with the simplest case of a simultaneous-move game with full information about payoffs, where the only uncertainty are opponents' action. After that, we extend this setting to more general cases, e.g. simultaneous-move games with incomplete information.

Simplest case 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 either a finite or an infinite number of rounds. A particular round is referred to as $t \in \mathbb{N}$. In every round t, every player i simultaneously chooses a specific action $a_t^i \in A^i$ out of a set of actions A^i, which can be either a continuous or a discrete action space.

There is a payoff mapping of the form $u^i : A^i \times A^{-i} \to \mathbb{R}$, giving von Neumann-Morgenstern utility u_t^i to each player i in every period t. At the beginning of period t, player i does not know

[5]Learning rules such as reinforcement learning (Roth and Erev, 1998) and experience-weighted attraction (Camerer and Ho, 1999) require a discrete action space. While these models can still be applied to continuous decision problems by discretization of the space, the bin size is an arbitrary decision and introduces a dilemma. Small bin sizes have two disadvantages: first, computations are slowed down and have been found not to converge. Second, because only a small part of the action space is "reinforced" after every feedback, even choices that are close to previous choices but do not coincide with them may be penalized by model selection criteria. On the other hand, a big bin size implies an imprecise approximation and imprecise predictions. See Arifovic and Ledyard (2004) for more details. Moreover, Hopkins (2002) shows that reinforcement and fictitious play, being a particular Bayesian model, "are far more similar than were thought."

the opponents' actions a_t^{-i}, so that she must form beliefs about them. The expected utility is given by $\mathbb{E}u^i : A^i \times \Delta(A^{-i}) \to \mathbb{R}$, where Δ is the set of all possible probability distributions. $\Delta(A^{-i})$ thus denotes the set of probability distributions over opponents' play. The (joint) probability of i's opponents' play in period t is given by a probability distribution $\sigma_t^{-i} \in \Delta(A^{-i})$. σ_t^{-i} is a probability density if A^{-i} is continuous and a probability mass function if A^{-i} is discrete.

Because $-i$ moves simultaneously to i, σ_t^{-i} is unknown to player i, so that she has a subjective probability distribution over opponents' play in t, denoted $\hat{\sigma}_t^{-i}$, which may be different from the objective probability distribution σ_t^{-i}. While $\hat{\sigma}_t^{-i}$ can be allowed to vary over time through a deterministic process, in the simplest case it is stationary like in the fictitious play model (Brown, 1951). Thus, solely for the purpose of notational convenience, we write $\hat{\sigma}^{-i}$.

Let i's subjective distribution $\hat{\sigma}^{-i} : \Theta \to \mathbb{R}^+$ where $\theta^i \in \Theta$ ($\Theta \subset \mathbb{R}$) is an exogenous parameter that defines that distribution. Thompson Sampling, like fictitious play, assumes that agents are uncertain about which θ^i best describes opponents' play and they update it based on past play, using Bayes' rule. Denote the prior of player i at the beginning of period t over θ^i by $D_{t-1}^i(\theta^i)$. Suppose player i observes opponents' play in t to be a_t^{-i}, then at the end of period t, she updates her prior by Bayes' rule in order to obtain the posterior:

$$D_t^i(\theta^i) = \frac{\hat{\sigma}^{-i}(a_t^{-i}|\theta^i) D_{t-1}^i(\theta^i)}{\hat{\sigma}^{-i}(a_t^{-i})} \tag{1}$$

The next decision player i faces is the choice of her action a_{t+1}^i in period $t+1$. A widely used assumption for decision-making in game theory is "rationalizability", being a weaker assumption than Nash equilibrium and only requiring that agents best-respond to any belief. In contrast to Nash equilibrium, this belief may not be objectively correct.

However, a best response requires laying out the complete game tree, which is computationally infeasible beyond setups with few rounds. (See e.g. Woodford (2018).) Setups in which learning by agents is analyzed usually contain many rounds. Thus, following a large literature in behavioral game theory, it is assumed that players give an asymptotically myopic best response, so that they choose the action that maximizes their immediate payoff. (See Fudenberg and Kreps (1993) for a discussion on myopia.)

Standard Bayesian learning corresponds to the case where $a_{t+1}^i = \arg\max \mathbb{E}_t(u_{t+1}^i|\hat{\sigma}^{-i})$. This is where Thompson Sampling differs from pure Bayesian learning. The myopic best response of

pure Bayesians implies (i) that player i makes optimal use of the posterior given her beliefs, so that her estimate of θ^i is $\mathbb{E}_t^i(\theta^{-i})$. This requires player i to calculate an integral or a weighted sum of all possible outcomes, using the entire posterior. Therefore, (ii) play of player i is deterministic.

Thompson Sampling represents a more coarse version that only makes use of a part of the posterior.[6] Specifically, Thompson Sampling corresponds to the case where each $a_{t+1}^i \in A^i$ is chosen according to the *probability* of maximizing $\mathbb{E}_t(u^i|\hat{\sigma}^{-i})$. This is equivalent to making a random draw from the posterior.[7] (See e.g. Chapelle and Li (2011).) Hence player i's estimate of θ^{-i} is a random draw $\tilde{\theta}_{t+1}^i \sim D_t(\theta^i)$, implying that player i's decision can be viewed as stochastic from an analyst's point of view. Instead of a best response, player i gives a *conditional* best response based on the random draw. Conditionally on the random draw $\tilde{\theta}_{t+1}^i$, player i's myopic best response would be: $a_{t+1}^i = \arg\max \mathbb{E}(u_{t+1}^i|\hat{\sigma}^{-i}(\tilde{\theta}_{t+1}^i))$.

Thompson Sampling thus consists of three steps: first, the player updates her subjective probability distribution in a Bayesian way; second, the player makes a random draw from the posterior; third, conditionally on that random draw, the player applies a (myopic) best response.

Given that both Thompson Sampling and Bayesian learning assume that actions are myopic best responses to some belief, one can say that both concepts are consistent with myopic rationalizability.

More general case The previous subsection describes Thompson Sampling under particular assumptions of the information set. This section relaxes these assumptions and describes Thompson Sampling for arbitrary information sets.

A particularly realistic case is that agents may have some uncertainty about payoff mappings. Thus, a payoff-relevant state $\omega_t \in \Omega$ can be introduced, being chosen by "nature" at the beginning of round t and being disclosed to a subset of players. The payoff function is then a mapping $u^i : A^i \times \Omega \times A^{-i} \to \mathbb{R}$. One can partition the domain of the payoff function into the domain of knowns, K^i, and the domain of unknowns, Γ^i. To apply Thompson Sampling, one must assume

[6] Previous papers introduced calculation costs into agents' decision-making such as Evans and Ramey (1992). Alaoui and Penta (2016) relate calculation cost to limited reasoning as described by "level k."

[7] The number of random draws could be an exogenous variable over which the mean is calculated. One draw is chosen for several reasons: first, the interpretation of probability matching would no longer hold. Second, if the number of draws is small, one would have to calculate the convolution (distribution of the sum of random variables) of D_t, which is intractable for a wide range of distributions. If the number of draws is large, the resulting distribution converges to a normal distribution by the Central Limit Theorem. Denoting the number of random draws by ν, the variance of the asymptotic distribution is given by $\frac{1}{\nu}\sigma^2$. This is undesirable for two reasons: (i) ν and σ^2 would hardly be identifiable; (ii) this can be summarized by one parameter, redefined as σ^2. In this light the empirical application in section 4 can be interpreted as agents making a large number of draws.

that the domain of unknowns is non-empty. For example, if the player knows only her own action but is neither informed about the state ω nor about opponents' actions a_t^{-i}, then $K^i = A^i$ and $\Gamma^i = \Omega \times A^{-i}$. Let $\sigma_t^{\gamma^i} \in \Delta(\Gamma^i)$ be the objective joint distribution over the unknowns. Denote player i's subjective distribution by $\hat{\sigma}_t^{\gamma^i}$ and let θ^i be the (vector of) exogenous parameters defining the distribution. Furthermore, the player receives observable signals, generated by a function $Z^i : A^i \times \Omega \times A^{-i} \to \mathbb{R}$. Denote the vector summarizing the available signals for player i at the end of period t by z_t^i. For example, if the payoff u_t^i is observable, then z_t^i may contain u_t^i.[8]

Thompson Sampling implies applying the three steps outlined above. First, the player updates her subjective probability distribution using Bayes' rule and the new information z_t^i:

$$D_t(\theta^i) = \frac{\hat{\sigma}^i(z_t^i|\theta^i)D_{t-1}^i(\theta^i)}{\hat{\sigma}^{\gamma^i}(z_t^i)} \tag{2}$$

Second, player i makes a random draw from the posterior: $\tilde{\theta}_{t+1}^i \sim D_t(\theta^i)$. Third, player i applies a best response conditional on the random draw. In case the player is myopic and only maximizes immediate payoffs, she would choose: $a_{t+1}^i = \arg\max \mathbb{E}(u_{t+1}^i|\hat{\sigma}^{\gamma^i}(\tilde{\theta}_{t+1}^i))$.

2.1 Relation to the previous literature

This section explores the precise relationship between Thompson Sampling and previous models of decision-making. Table 1 gives a non-exhaustive overview of some frequently used models in the literature. This table shows that those models can be classified along two dimensions: the manner of belief formation and how people select their actions. Since the main novelty of Thompson Sampling is how agents use their belief distribution to select their actions, the discussion is organized along action selection. Table 1 shows two main ways of action selection: optimal and through a random utility model. The relationship between Thompson Sampling and the random utility model is discussed below.

Optimal action selection The novelty of Thompson Sampling is how agents use their belief (distribution). The assumption that agents update beliefs in a Bayesian way has been used by previous approaches such as fictitious play (Brown, 1951) or internal rationality (Adam and Marcet, 2011). Similarly to equilibrium models such as Nash equilibrium or rational expectations,

[8]Section 4 provides an example, where u_t^i is a deterministic function of the actual signal so that u_t^i does not contain any additional information.

these models are deterministic, assuming that agents select the optimal action conditional on their beliefs. Randomness in these models does not occur other than due to deterministic (and optimal) responses to exogenous randomness in fundamentals. Thompson Sampling assumes that agents select the optimal action conditionally on a random draw from the posterior. This renders Thompson Sampling a stochastic models. Furthermore, since the Bayesian posterior is constructed from past data of the environment, the randomness evolves endogenously.

Random utility models No matter how beliefs are formed, a model can be augmented by introducing an exogenous random component determining agents' choice (second row center). A popular way of doing so is the random utility model. The *generalized random utility* model[9] is commonly defined as a probability measure D on a set of utility functions. (Gul and Pesendorfer, 2006)

The simplest case, being referred to as *standard random utility model* here, has been assumed to be additive in two terms: $U_t^i = V_t^i + \epsilon_t^i$, where V_t^i is a deterministic part that varies over time and ϵ_t^i is an i.i.d. random shock being drawn from a *static* distribution $D(\epsilon^i)$, where D is often specified as type I extreme value.[10] In this sense, the standard random utility model can be considered a static model. There have been many proposals regarding the specification of V_t^i in the learning literature. Quantal response equilibrium (QRE) (McKelvey and Palfrey, 1995) is a specific hypothesis, purporting that this part is calculated rationally in the sense that $V^i(.)$ 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 (EWA) (Camerer and Ho, 1999), assume a boundedly rational way of calculating $V^i(.)$, not using the assumption of equilibrium beliefs.

While i.i.d. random errors from a static distribution can be considered a convenient statistical assumption, researchers have provided different motivations for relaxing this assumption.[11] One important violation of i.i.d. is heteroskedasticity of the perturbations. If heteroskedasticity

[9]There is no clear consensus in the literature on the use of the term random utility model. Thus, the term generalized random utility model is borrowed from Walker and Ben-Akiva (2002), encompassing all models where the definition given by Gul and Pesendorfer (2006) applies. This includes but is not limited to the standard random utility model defined below, models with flexible disturbances like random parameter models and latent class models.

[10]An alternative specification in the previous learning literature is a power probability specification. For a discussion of the relative merits and drawbacks of the power and the logit specifications, see Camerer and Ho (1999).

[11]See e.g. Louviere and Eagle (2006), Train (2009) or Fiebig et al. (2010) for an overview.

is ignored, it can have severe consequences such as inconsistent parameter estimates and biased forecasts. (Louviere and Eagle, 2006) Explicitly modeling this heteroskedasticity only solves the problem if it is correctly specified. Otherwise, augmenting the model for heteroskedasticity comes at the additional expense of increased complexity. This is why authors such as Louviere and Eagle (2006) propose that "without theory to suggest how components of variance differ by individuals, markets, contexts, experiments, etc, adding higher moments to choice models is probably a bad idea" and "[a] better way forward is to develop theory and methods to capture variability differences."

Absence of a compelling theory about how the moments of the distribution of utility may evolve over time motivated Frick et al. (2017) to provide the first analysis of the fully general, non-parametric form of a dynamic version of the random utility model where the distribution D_t^i is time-varying. They provide the axiomatic characterization, also considering the case where D_t^i depends on past choices.

Another questionable feature of a random utility model with i.i.d. shocks has been found by Apesteguia and Ballester (2017). They show that every i.i.d. random utility model implies non-monotonic risk preferences, i.e. there is a range of parameters where the probability of making a riskier choice increases in the level of risk aversion. They show that monotonicity is, however, preserved in random parameter models, in which the randomness is introduced through stochastic parameters in the utility function. This relaxes the strong i.i.d. assumption of the utility perturbations and introduces correlation over actions.

Thompson Sampling is related to Frick et al. (2017), as it also embarks on the idea of an endogenous distribution D_t^i. It proposes a specific theory that retains computational simplicity and tractability by linking the randomness in choice to the Bayesian posterior, summarizing information acquired about the environment. Moreover, it preserves monotonic risk preferences, as it belongs to the random parameter class of models: the randomness comes from drawing the parameter θ^i, which enters the subjective payoff consideration for every possible action of the player and thus introduces correlated perturbations.

2.2 Evaluating Thompson Sampling

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

Action selection		Beliefs: Equilibrium	Beliefs: Bayesian	Beliefs: Other (e.g. learning)
	Optimal	Nash equilibrium Rational expectations Perfect foresight eq.	fictitious play internal rationality	least square learning constant gain learning
Generalized Random Utility	**Exogenous trembles**	*quantal response equilibrium*	*Bayesian learning with exogenous shocks*	reinforcement learning EWA heuristic switching
Generalized Random Utility	**Optimal response to sample from posterior**	N/A	*Thompson Sampling*	N/A

Table 1: Thompson Sampling in comparison to previously used approaches

model: first, just as Thompson Sampling, the benchmark model should also allow for stochasticity. Second, 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. 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 available 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, which is i.i.d. over actions in *all* known empirical applications of QRE due to practical limitations. (Goeree et al., 2016, p. 48) Thompson Sampling is a simple way to introduce correlated shocks through random draws from the posterior. While the distribution of the errors is arbitrary, it is usually taken as 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 for Thompson Sampling and a logit error structure, labeled

as "Bayesian 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 estimates expected payoff using Bayesian learning, being the same way of updating as 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_t^{-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 *correct* probability distribution[12] of others σ^{-i}

$$V^i(a_t^i = a^{ij}) = \mathbb{E}t - 1(u_t^i | a_t^i = a^{ij}, \sigma^{-i}) \tag{3}$$

Perhaps the most common distribution, particularly in the literature applying QRE to a continuous action space (see Goeree and Holt (2005) for an overview), is a type I extreme value distribution, resulting in the logit version of QRE. Less common is a multinomial probit specification due to practical limitations. (See Cameron and Triverdi (2005) and Train (2009) for a discussion.) Furthermore, the results are generally very similar between logit and probit. (See e.g. Long (1997), p. 83, Greene (2002), p. 667, or Gill (2001), p.33.)

Assuming a type I extreme value distribution, the probability of choosing any action a^{ij} is given by the logit model:

$$P(a_t^i = a^{ij}) = P(U_t^i(a_t^i = a^{ij}) \geq U_t^i(a_t^i = a^{ik}), \forall k \in J_i) = \frac{\exp(\lambda \mathbb{E}(u_t^i | a_t^i = a^{ij}, \sigma^{-i}))}{\sum_{k=1}^{J_i} \exp(\lambda \mathbb{E}(u_{t+s}^i | a_t^i = a^{ik}, \sigma^{-i}))} \tag{4}$$

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 $\hat{\sigma}^{-i}$ or a perceived law of motion (PLM) about how the (market) outcomes are generated that may not coincide with the actual law

[12] As every player i forms consistent beliefs and rationally calculates the reward, the past does not play any role.

of motion (ALM).[13] Those non-equilibrium beliefs $\hat{\sigma}^{-i}$ may contain states θ^i that are unknown to agent i and about which she forms beliefs, $D_t^i(\theta^i)$. It is assumed that as new information becomes available, the decision-maker updates $D_t^i(\theta^i)$ in a Bayesian way. Like in QRE, the decision-maker does not best respond to those beliefs but instead trembles when choosing the action. It has been shown that fictitious play corresponds to Bayesian learning with Dirichlet priors, which is the multivariate generalization of the beta distribution. (DeGroot, 1970) Thus, this class of models comprises *stochastic fictitious play*, which has frequently been used in the previous behavioral game theory literature. (See e.g. Fudenberg and Kreps (1993), Fudenberg and Levine (1998), Cheung and Friedman (1997), Goeree and Holt (1999).)

Given Bayesian beliefs, an agent calculates the expected reward

$$V^i(a_t^i = a^{ij}) = \mathbb{E}_{t-1}(u_t^i | a_t^i = a^{ij}, \hat{\sigma}^{-i}(\theta^i)) \tag{5}$$

For the same reasons as stated above for the application of QRE, a type I extreme value distribution is specified, so that the probability distribution over the action space is given by the logit model:

$$P(a_t^i = a^{ij}) = P(U_t^i(a_t^i = a^{ij}) \geq U_t^i(a_t^i = a^{ik}), \forall k \in J_i) = \frac{\exp(\lambda \mathbb{E}(u_t^i | a_t^i = a^{ij}, \hat{\sigma}^{-i}))}{\sum_{k=1}^{J_i} \exp(\lambda \mathbb{E}(u_{t+s}^i | a_t^i = a^{ik}, \hat{\sigma}^{-i}))} \tag{6}$$

This is why this model is referred to as the *Bayesian logit*.

3 Application to 2x2 games

3.1 Dataset

10 constant-sum games from Erev et al. (2007) are taken (payoffs shown in table 2), in which each subject played 500 periods of only one of these games against the same opponent. Every game is played by nine pairs, so that this dataset contains altogether 90,000 observations. The numbers in each cell represent the probabilities that players win a fixed lottery prize υ, set to \$0.04, on each trial. For instance, if ROW plays T and COL plays L, player 1 will win υ with the specified probability P_1^{ROW}, while player 2 will win υ 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

[13] PLM and ALM is the terminology used by the macroeconomic learning literature. See for example Evans and Honkapohja (2001).

the probabilities in the payoff matrix of 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 prize. However, players were not informed whether the opponent received the lottery prize or not.

Game		Payoff matrix		Game		Payoff matrix	
		L	R			L	R
1	T	.77	.35	6	T	.46	.54
	B	.08	.48		B	.61	.23
		L	R			L	R
2	T	.73	.74	7	T	.89	.53
	B	.87	.20		B	.82	.92
		L	R			L	R
3	T	.63	.08	8	T	.88	.38
	B	.01	.17		B	.40	.55
		L	R			L	R
4	T	.55	.75	9	T	.40	.76
	B	.73	.60		B	.91	.23
		L	R			L	R
5	T	.50	.64	10	T	.69	.05
	B	.93	.40		B	.13	.33

Table 2: Games in Erev et al. (2007)

The dynamic patterns of four of those games are shown in figure 1. It is of particular interest whether the noise patterns are similar over time and 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. (p-values > 0.05)[14][15]

Different degrees of volatility across games are expected due to different Nash equilibria. For example, if the mixed strategy Nash equilibrium is 50-50, then equilibrium play, involving play of both actions with equal frequencies, has a greater variance than for a Nash equilibrium 80-20, where one action is played much more often than the other.

[14]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).

[15]A potential concern is that there are only nine independent observations for each game, which leads to low power of statistical tests. Yet, the results are robust to pooling over games, since sign tests based on 90 independent observations obtain insignificant results for both player 1 (p-value: 0.5203) and player 2 (p-value: 0.3912).

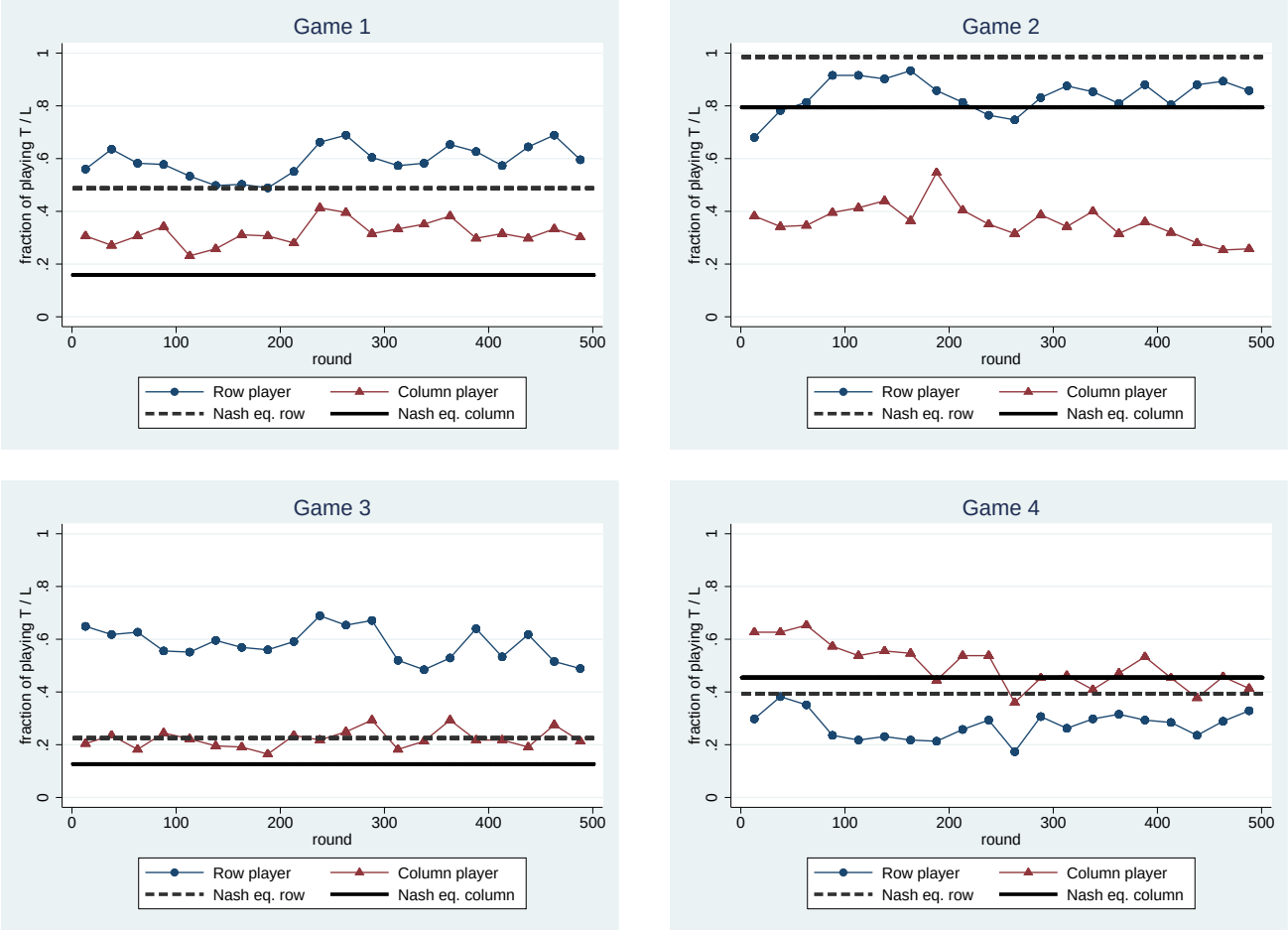

Figure 1: Constant-sum one-stage 2x2 games with full information played over 500 rounds from Erev et al. (2007): the proportion of T-choices for ROW and L-choices for COL is grouped in blocks of 25 periods

3.2 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 τ rounds, indexed by $t = 1, 2, ...\tau$. The probability distributions of play are $\sigma_t^i = (p_t^i(a^{i,1}), p_t^i(a^{i,2})) = (p_t^i(a^{i,1}), 1 - p_t^i(a^{i,2}))$, with $p_t^i(a^{i,1})$ denoting the objective probability that player i plays $a^{i,1}$ in period t. The payoff u_t^i is given by a von-Neumann-Morgenstern payoff matrix in table 3.[16]

[16] 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 v with probability P_1^R, while COL does with probability P_1^C.

		COLUMN	
		$a^{COL,1}$	$a^{COL,2}$
ROW	$a^{ROW,1}$	P_1^{ROW}, P_1^{COL}	P_2^{ROW}, P_2^{COL}
	$a^{ROW,2}$	P_3^{ROW}, P_3^{COL}	P_4^{ROW}, P_4^{COL}

Table 3: Exemplary payoff matrix

Under full information about the opponent's past play and the payoffs, any player i's only unknown at time t is the opponent's contemporaneous action a_t^{-i}. However, she observes all past actions directly so that they constitute the observed signals $z^{*,i}(a_t^{-i}) = a_t^{-i}$.

3.2.1 Beliefs about the other player's action

I closely follow the literature on fictitious play in specifying agents' beliefs. By construction of the setup, the players move simultaneously in any period t. Thus, player i does not know the objective probability distribution of her opponent's play σ_t^{-i}. Hence, she must guess a subjective likelihood $\hat{\sigma}_t^{-i}$ with parameters θ^i. While many candidates for $\hat{\sigma}_t^{-i}$ would be possible, for this application I specify it as a Bernouilli distribution, denoted Ber. Like in fictitious play, this implies a stationary distribution, so that players do not take into account that other players learn over time.[17] There is one unknown parameter $\theta^i = p^{-i}(a^{-i,1})$, corresponding to the probability that player $-i$ plays $a^{-i,1}$, which defines a Bernouilli distribution $Ber(p^{-i}(a^{-i,1}))$.

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

$$p^{-i}(a^{-i,1}) \sim \mathfrak{B}(\alpha_{t-1}^{-i}, \beta_{t-1}^{-i}) \tag{7}$$

where α_{t-1}^{-i} represents the number of previous trials in which the opponent -i indeed played $a^{-i,1}$. β_{t-1}^{-i} represents the number of previous trials in which the opponent played $a^{-i,2}$.

A beta distribution is chosen for several reasons: first, it is the conjugate prior of the Bernouilli distribution, meaning that under a Bernouilli distributed outcome the posterior distribution is of

[17] 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).) See also Fudenberg and Levine (1998) for a discussion.

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.

3.2.2 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

$$\mathbb{E}^R_{t-1} p^{-i}(a^{-i,1}) = \frac{\alpha^{-i}_{t-1}}{\alpha^{-i}_{t-1} + \beta^{-i}_{t-1}} \tag{8}$$

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 (7), 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 3: With $\tilde{p}^{COL}_t(L)$ as her estimate for the column player to play L, the row player's discounted expected payoffs are, for instance, if she plays L:

$$\mathbb{E}(u^{ROW}_t | a^{ROW}_t = L, Ber(\tilde{p}^{COL}_t(L))) = P^{ROW}_1 \cdot \tilde{p}^{COL}_t(L)_t + P^{ROW}_2 \cdot (1 - \tilde{p}^{COL}_t(L)) \tag{9}$$

Hence, it is easy to see that the row player plays L if $\mathbb{E}(u^{ROW}_t | a^{ROW}_t = L, Ber(\tilde{p}^{COL}_t(L))) > \mathbb{E}(u^{ROW}_t | a^{ROW}_t = R, Ber(\tilde{p}^{COL}_t(L)))$ and R if $\mathbb{E}(u^{ROW}_t | a^{ROW}_t = R, Ber(\tilde{p}^{COL}_t(L))) > \mathbb{E}(u^{ROW}_t | a^{ROW}_t = L, Ber(\tilde{p}^{COL}_t(L)))$.

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

$$p^{ROW}_t(L) = \begin{cases} 1 - I_p(a^{COL}_{t-1}, b^{COL}_{t-1}) & \text{if } P^{ROW}_1 - P^R_2 - P^{ROW}_3 + P^{ROW}_4 > 0 \\ I_p(a^{COL}_{t-1}, b^{COL}_{t-1}) & \text{otherwise} \end{cases} \tag{10}$$

where I_p is the "regularized incomplete beta function", the c.d.f. of the beta function.

3.2.3 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 $D_{t-1}(p^{-i}(a^{-i,1}))$.

A specific property of the beta distribution is that Bayesian updating, as generally formulated by equation (1), implies adding 1 to α_{t-1}^{-i}, if she observes the opponent playing $a^{-i,1}$ in period t and adding 1 to β_{t-1}^{-i}, if she observes the opponent playing $a^{-i,2}$ in period t. Hence:

$$\alpha_t^{-i} = \begin{cases} \alpha_{t-1}^{-i} + 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \alpha_{t-1}^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{11}$$

$$\beta_t^{-i} = \begin{cases} \beta_{t-1}^{-i} + 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \beta_{t-1}^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{12}$$

The beta distribution $\mathcal{B}(\alpha_t^{-i}, \beta_t^{-i})$ constitutes player i's posterior belief $D_t^i(p^{-i}(a^{-i,1}))$.

Generalization of Bayesian updating The unknown parameters are the priors at the beginning of the game α_0^{-i} and β_0^{-i}. Estimating those parameters under pure Bayesian belief updating yields extremely high parameter estimates. A wide range of applications not only in the decision-making but also in the behavioral game theory literature explores the hypothesis of players having biased perceptions when updating. (Fudenberg and Levine, 1998; Roth and Erev, 1998; Camerer and Ho, 1999) The approach taken here follows Goeree et al. (2007), who develop a generalization of Bayesian updating, and Moreno and Rosokha (2016), who generalize their framework to a setting with many time periods.

Bayes' rule as given by equation (1) can more generally be written as

$$D_t^i(\theta^i) = \frac{(\hat{\sigma}^{-i}(a_t^{-i}|\theta^i)))^{\xi^t} D_{t-1}^i(\theta^i)}{(\hat{\sigma}^{-i}(a_t^{-i}))^{\xi^t}} \tag{13}$$

The appeal of this specification is its flexibility given by the parameter ξ, which captures the perceived number of signals. Pure Bayesian learning is nested by setting $\xi = 1$. If the agent after observing the next signal acts as if she observed two signals, then $\xi = 2$. Values of $\xi > 0$ can be interpreted as limited memory, since agents pay more attention to more recent signals. Conversely, values of $\xi < 0$ can be interpreted as underweighting of the signal or "conservatism

bias." To distinguish between old and more recent periods, following Moreno and Rosokha (2016), the weight of the signal in a period t is ξ^t, meaning that each new signal has ξ times the weight of the previous signal.

This implies an updating rule of

$$\alpha_t^{-i} = \begin{cases} \alpha_{t-1}^{-i} + \xi^t \cdot 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \alpha_t^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{14}$$

$$\beta_t^{-i} = \begin{cases} \beta_{t-1}^{-i} + \xi^t \cdot 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \beta_{t-1}^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{15}$$

3.2.4 Initial priors

A remaining question is how to specify the initial priors. Equations (38) and (39) highlight that beliefs must be initialized by specifying α_0^{-i} and β_0^{-i}. Those parameters reflect the numbers trials, in which the opponent played $a^{-i,1}$ or $a^{-i,2}$ 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 $a^{-i,1}$ as opposed to $a^{-i,2}$, which is reflected by the ratio $\frac{\alpha_0^{-i}}{\beta_0^{-i}}$ as well as the sample size of this prior "hypothetical sample" $N \equiv \alpha_0^{-i} + \beta_0^{-i}$. A high magnitude of N 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 N is low, the 1's that are added during the play carry more weight.

There is no consensus in the literature on how to initialize priors. (See e.g. Williamson (2010) for a discussion.) Thus, in a first step the priors have been estimated for Thompson Sampling and the Bayesian logit together with the exogenous parameters. Since different information was given to the subjects in each of the 10 games (e.g. games differ in the payoff matrices), it is plausible that the priors differ over the games. 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 of the same type.

The data display stark differences in initial play. It would thus be incorrect to use the initial conditions of one game to predict the dynamics of another game. However, if the variation in the priors is not understood, this will present an obstacle to forecasting the dynamics of new games ex-ante. Thus, the estimates of the prior parameters have been investigated for regularities.

For Thompson Sampling and the Bayesian logit, the mean estimates $\hat{p}_0^{-i}(L) = \frac{\alpha_0^{-i}}{\beta_0^{-i}+\alpha_0^{-i}}$ have been found to be relatively close to the Nash equilibria of every game.[18] Priors being close to the mixed strategy equilibrium are sensible in the context of Thompson Sampling. The reason is that they induce about equal payoff associated with each of the two actions so that the player's decisions more likely look like a mixed strategy.

The prior probabilities $\hat{p}_0^{-i}(L) = \frac{\alpha_0^{-i}}{\beta_0^{-i}+\alpha_0^{-i}}$ thus have been restricted to the Nash equilibria, so that the only unknown parameter is the "prior" sample size, being denoted by N.

3.3 Empirical evaluation

3.3.1 Methodology

The initial conditions (or priors) are estimated together with the rationality parameter λ. 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 reasonable, as the environment to which the subjects are exposed 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, fitting a model in-sample carries the peril of overfitting (see e.g. Leamer (1978)), meaning that the parameter estimates are driven by noise in the calibration dataset. This is a particular concern if models with different numbers of parameters are compared, since models with more exogenous parameters might have an advantage in fitting the data. A popular way to guard against overfitting is the adoption of information criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). (Akaike, 1974; Schwarz, 1978) However, it has been shown that the model selected by the BIC does not necessarily have the best out-of-sample predictability (see e.g. Hansen (2010)). In fact, the ultimate test of a model is considered to be (pseudo) out-of-sample forecasting. (Chatfield, 1996; Stock and Watson,2015, p. 613) This is a common methodology both in microeconomics (e.g. Camerer and Ho (1999)) and macroeconomics (see Clark and McCracken

[18]For Thompson sampling the median deviations from Nash equilibrium for the row players ($-2.53 \cdot 10^{-6}$) and the column players ($-1.56 \cdot 10^{-7}$) were insignificantly different from zero (p-value Wilcoxon signed-rank test: 0.6250 for ROW; 0.8457 for COL.) For the Bayesian logit the median deviations from Nash equilibrium for the row players (-0.0426) and the column players (0.0061) were also insignificantly different from zero. (p-values of Wilcoxon signed-rank tests: 0.2754 for ROW; 0.5566 for COL). Separate statistical tests need to be conducted for the row and the column player, as otherwise the independence assumption that is required for the Wilcoxon signed-rank test would be violated.

(2013)). Moreover, models with more free parameters do not necessarily provide better out-of-sample forecasts than models with fewer parameters.

To test the models' predictive performance, 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 samples usually render conclusions about model evaluation more robust.

To assess the in-sample fit, I use the in-sample likelihood and two measures penalizing models with extra exogenous parameters: the AIC and the BIC. For out-of-sample forecasting, the BIC is commonly not used, since models with more free parameters generally do not have any advantage when forecasting out-of-sample. Thus, I focus on the log-likelihood (LL) as a loss function for the validation sample:

$$LL = \sum_{t=1}^{T} \sum_{i=1}^{n_v} \ln(f(a_t^i)) \tag{16}$$

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.

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. Cross-validation with 10-20 breaking points has been recommended, since too many breaking points lead to a high variance and thus inefficient estimates. (Kohavi, 1995) 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.

This procedure also has the advantage that it provides a jackknife estimate of the standard error. Since 10-fold cross-validation provides 10 different parameter estimates, standard errors can be estimated by calculating the sample standard deviation of the distribution of parameters obtained.

QRE Applying QRE is straightforward. The logit equations constitute two equations with two unknown probabilities for every λ. Solving for these probabilities can be included in any search algorithm, finding the λ-parameter, which maximizes the log-likelihood of the training sample.

TS Thompson Sampling has one free parameter: N. The restrictions $\alpha_0^{-i} = p^{-i,*} \cdot N$ and $\beta_0^{-i} = (1 - p^{-i,*}) \cdot N$ 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. Note that the one exogenous parameters for Thompson Sampling is introduced to pin down the initial prior. For the learning process itself, no exogenous parameter is necessary.

Logit The logit approach uses the same error structure as QRE (with one exogenous parameter) but Bayesian learning. To be consistent, Bayesian learning has been specified the same way as in Thompson Sampling with a Bernouilli likelihood and a Beta prior. This results in two free parameters: N the "prior" sample size, the same exogenous parameter as in Thompson Sampling, as well as the λ-parameter, capturing the precision (inverse variance) of the action selection.

3.3.2 Estimation results

Model comparison Regarding in-sample fit, Thompson Sampling, Bayesian learning and QRE all perform better than Nash equilibrium and a random uniform. These results are robust to penalizing them according to their number of exogenous parameters. Notably, Thompson Sampling despite having one parameter less than Bayesian learning Thompson Sampling provides the better in-sample fit. Amongst all models, the best in-sample fit is provided by QRE. This is unsurprising, given some results in the previous literature. Haile et al. (2008) show that QRE is not falsifiable in any normal form game, if the i.i.d. assumption of the perturbations is slightly relaxed. However, even imposing the i.i.d. assumption can explain a large set of outcomes. Furthermore, the logit version of QRE does not rule out much in 2x2 games. However, Haile et al. (2008) argue that

in games with more than two actions the results that a logit-QRE can rationalize is considerably more limited. Hence, the test on a dataset with a large action space presented later is particularly important.

	TS	**Logit**	**QRE**	**Random**	**Nash**
TS	-	-	-	-	-
Logit	Tie (0.0588)	-	-	-	-
QRE	QRE (0.0069)	QRE (0.0093)	-	-	-
Random	TS (0.0051)	Logit (0.0069)	QRE (0.0051)	-	-
Nash	Tie (0.2041)	Tie (0.1676)	QRE (0.0124)	Tie (0.4473)	-

Table 4: 2x2 games: Preferred model for out-of-sample forecasting by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

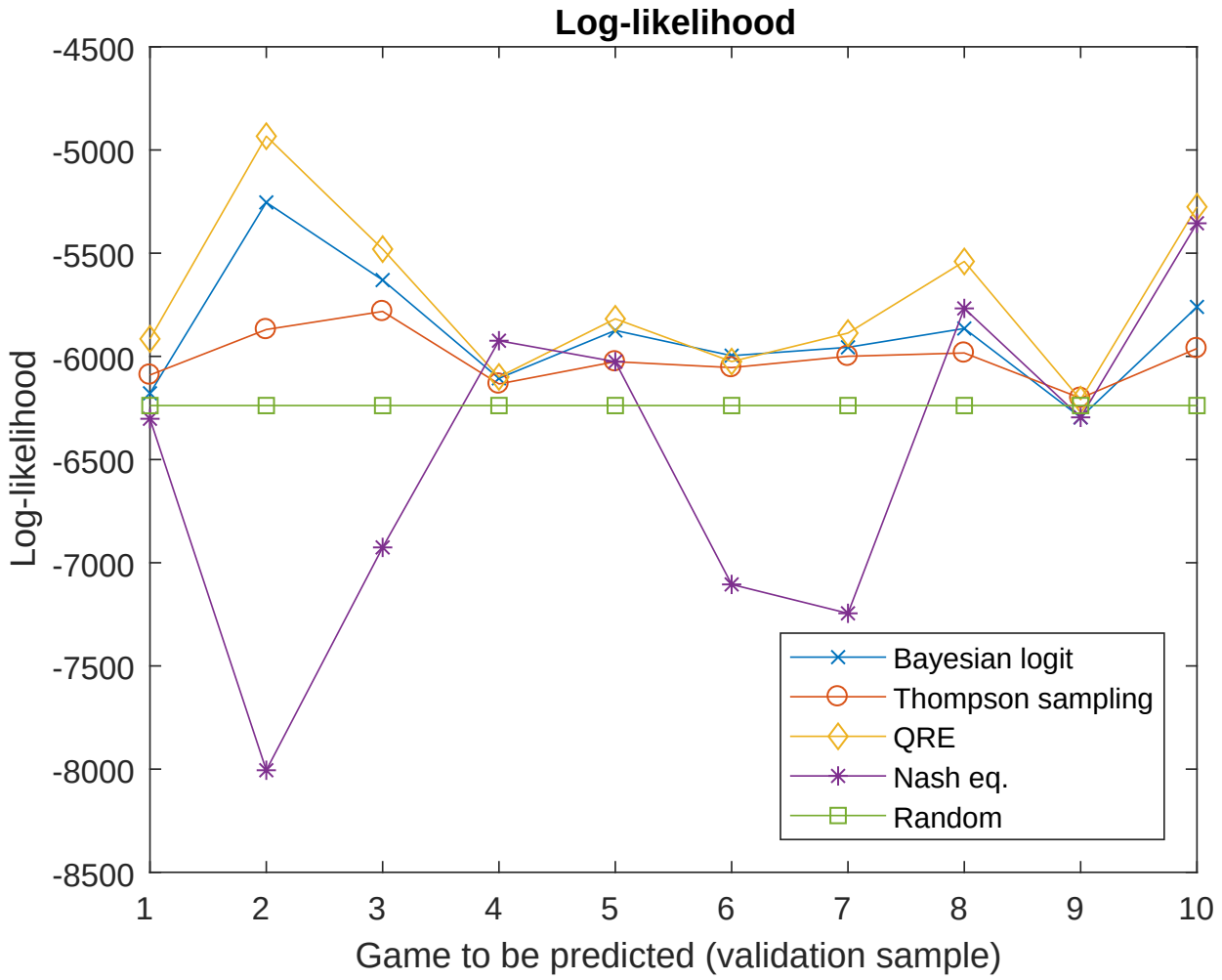

Figure 2: 2x2 games: Likelihoods of validation samples

Turning to out-of-sample predictions, table 4 reports the results of Wilcoxon signed-rank tests that compare the out-of-sample performance of the models pairwise and figure 2 shows the likelihoods obtained from the cross-validation procedure. Thompson Sampling, the Bayesian logit and

QRE all provide better forecasts than random decision-making (choosing each action with a 50 % probability) in nine out of ten games. Hence, signed-rank tests reject the hypothesis that a random uniform predicts observed behavior equally well as all three models examined. Furthermore, all three models predict better than Nash equilibrium. Random does better than Nash equilibrium in six out of ten games. Hence, there is no evidence that Nash equilibrium predicts behavior better than random. Yet, only QRE does significantly better than Nash equilibrium, while for the other two models there is no clear evidence that they outperform Nash equilibrium. The out-of-sample predictive ability of Thompson Sampling can be improved if the model is augmented by psychological theories as it has been done in the behavioral game theory literature before. (See Appendix .) If more parameters are added to Thompson Sampling it delivers approximately equal out-of-sample predictions to QRE. A better fit cannot necessarily be expected: as there was no heteroskedasticity detected in the dataset, Thompson Sampling does not have a comparative advantage here.

Parameters Table 5 reports the maximum likelihood parameters estimated as well as the jack-knife standard errors. The following observations stand out:

Observation 1. The priors for successes and failures TS take very high values, while they take low values for the Bayesian logit.

The magnitudes of α_0^{-i} and β_0^{-i} 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 α_0^{-i} and β_0^{-i} is low, the 1's that are added during the play carry more weight.

For TS, α_0^{-i} and β_0^{-i} 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 type I extreme value distribution used for the logit has heavy tails and thus easily predicts large shocks in the action space; conversely, TS postulates randomness in the beliefs so that small shocks around the knife-edge value of opponents'

probability of play suffice to create variation in players' action selection.

	Size of hypothetical prior sample	Rationality parameter
	$\underline{\mathbf{N}}$	$\underline{\lambda}$
Logit	4.04	3.50
	(1.15)	(0.16)
TS	432,597.88	-
	(40,824.77)	
QRE		6.31
		(0.26)

standard errors in parentheses

Table 5: MLE estimates

4 Application to expectation formation

Experiments have shown that, even in the presence of a unique equilibrium, learning dynamics on a continuous action space depend on the kind of feedback in the underlying system (Camerer 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 $p_t^{e,i}$ (corresponding to the actions) for the price realization in period t are inserted into a price adjustment equation:

$$p_t = c + b\frac{1}{n}\sum_{i=1}^{n} p_t^{e,i} + \epsilon_t \tag{17}$$

c, b represent exogenous parameters and $\epsilon_t \sim N(0, \frac{1}{4})$ represents a stochastic shock. Equation (17) (including the shock realizations ϵ_t) is never disclosed to the participants. In fact, this setup represents a beauty-contest game with an interior solution and stochastic shocks.

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.

It can easily be verified that this implies a unique fundamental $p^f = 60$ as well as a unique rational expectations equilibrium.[19] Participants are rewarded according to a quadratic distance equation:

$$u_t^i = \max\{0, 1300 - \frac{1300}{49}(p_t - p_t^{e,i})^2\} \tag{18}$$

The results of the different experimental groups are displayed in figures 3 and 4. For strategic substitutes, Heemeijer et al. (2009) find quick convergence to the fundamental, while for strategic complements, there is no evidence for convergence.[20] However, coordination in the strategic

[19] The rational expectations equilibrium is a constant, since the shock is only realized at the end of a period.

[20] 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.

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. Sign tests indicate that the variance in subjects' forecasts in the first 38 rounds (75 % of the sample) is higher than in the last 12 rounds (25 %) for both the strategic substitute (p-value: 0.0156) and complement treatment.[21] (p-value: 0.0078)

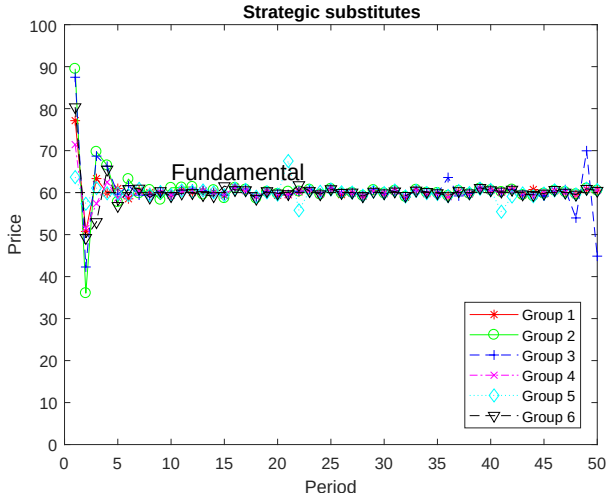

Figure 3: Strategic substitutability sessions from Heemeijer et al. (2009)

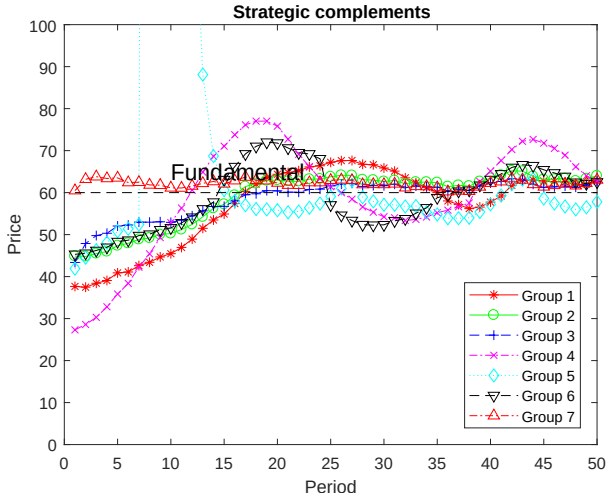

Figure 4: Strategic complementary sessions from Heemeijer et al. (2009)

4.2 Theory

Suppose agents do not know the price adjustment equation (17). 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.[22] 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:

$$p_t \sim N(p^*, \sigma^2) \tag{19}$$

$$\text{or alternatively: } p_t = p^* + \eta_t \text{ with } \eta_t \sim N(0, \sigma^2) \tag{20}$$

[21] This result is based on treating 1 single observation causing the big spike in group 3 period 49. Not excluding that observation would render the test for strategic substitutes insignificant. (p-value: 0.1094)

[22] This is the natural starting point, since any misperception about the underlying law of motion in the boundedly rational models, i.e. Thompson Sampling and Bayesian learning, would make it harder to disentangle whether differences to the equilibrium models come from learning or the misperception.

p^* is unknown and corresponds to the variable or state θ^i $(\forall i \in I)$ that needs to be learned. For technical simplicity, I assume that the variance σ^2 is known (or the player believes to know the variance).[23] η_t can been seen as the (perceived) stochastic part in determining p_t from the perspective of the subjects.

Consider a period t where each player i needs to forecast p_t given past data until period t-1.[24] Given (19), the optimal forecast is $\mathbb{E}_{t-1}p_t = p^*$. 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, the player has to form a prior belief about it in any period t. While this prior could in principle take any form, it is, for technical simplicity, assumed to be Gaussian $N(\bar{p}^*_{t-1}, \rho_{t-1})$. Suppose a new observation p_t just became available. With p_t, the prior can be updated in a Bayesian way to obtain the posterior $N(\bar{p}^*_t, \rho_t)$. The Bayesian update is given by the Kalman filter (Kalman, 1960). The Kalman filter uses a filtering equation of the form:

$$\bar{p}^*_t = \bar{p}^*_{t-1} + g_t(p_t - \bar{p}^*_{t-1}) \tag{21}$$

where g_t is the gain parameter.[25] 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+1}, and the posterior mean, $\bar{p}^*_t$:

$$\kappa_t \equiv \arg\min_{g_t} \mathbb{E}_t[\bar{p}^*_{t-1} + g_t(p_t - \bar{p}^*_{t-1}) - p_{t+1}]^2 = \frac{\rho_{t-1}}{\rho_{t-1} + \sigma^2} \tag{22}$$

Equation (22) implies that the findings documented 2x2 games that agents attach declining weights on new observations is already ingrained in Bayesian learning with a Gaussian prior. At the same time, equation (22) predicts that agents discount old observations.

The variance of $\bar{p}^*_t$, denoted by ρ_t, is obtained as

$$\rho_t = (\frac{1}{\rho_{t-1}} + \frac{1}{\sigma^2})^{-1} \tag{23}$$

[23]The variance σ^2 is not necessarily the same as the variance of the exogenous ϵ-shocks.

[24]p_t is yet unknown at the beginning of period t, as it depends on p^e_t, the forecast that has to be submitted in period t.

[25]Specification (21) is the cornerstone the adaptive learning class of models. Several proposals have been made regarding the specification of the gain g_t such as least square learning or constant gain learning. (See Evans and Honkapohja (2001) for an overview.) Constant gain learning has been shown to be an approximation to Kalman filtering. (Evans et al., 2010)

An optimizing Bayesian agent would forecast $\mathbb{E}_t p_{t+1} = \mathbb{E}_t p^* = \bar{p}_t^*$, which results in a deterministic model. However, an agent using Thompson Sampling proceeds through the three steps (Bayesian updating, random draw, myopic best response) in the theory section. To instantiate her belief about p^*, she makes a random draw $\tilde{p}_t^*$ from the posterior $N(\bar{p}_t^*, \rho_t)$. Conditionally on $\tilde{p}_t^*$, she chooses $p_{t+1}^{e,i}$ as to maximize her expected reward of the next period. Using (20), the expected reward can be written as:

$$\mathbb{E}_t(u_{t+1}^i | N(\bar{p}_t^*, \rho_t)) = \mathbb{E}_t[\max\{0, 1300 - \frac{1300}{49}(\tilde{p}_t^* + \eta_t - p_{t+1}^{e,i})^2\}] \tag{24}$$

which implies an optimal action of $p_{t+1}^{e,i} = \tilde{p}_t^*$. Once every individual has made her choice, the average price forecast in period $t+1$, $p_{t+1}^e \equiv \frac{1}{n}\sum_{i=1}^n p_{t+1}^{e,i}$, can be obtained and p_{t+1} can be calculated and announced to the players. The same steps are then repeated: p_{t+1} is then used to obtain a new posterior from which a random draw is made to obtain the forecast for period t+2 etc..

To uniquely determine the distributions that generate the individual beliefs, the prior mean $\bar{p}_0^*$, the prior variance ρ_0 and the variance of the perceived shock, σ^2, need to be calibrated. A question is whether one should allow for different priors like for the 2x2 games. The learning-to-forecast experiment deals with incomplete information, where subjects receive only qualitative instructions. The only quantitative instruction that is given in both treatments is: "The price(and your prediction)can never become negative and lies always between 0 and 100 euros in the first period." Thus, it is reasonable to assume that subjects have the same initial priors.

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 procedure 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 independent 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 the Bayesian logit on the continuous action 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.[26] I divide the space up into 100 equal intervals:[27] $[0,1),[1,2),...,[99,100]$. As exemplified below, the Bayesian logit corresponds to a particular case of adaptive learning with shocks. 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.[28] A preliminary question is whether λ 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 mid-point of each interval is used for payoff consideration, which is $p^{e,ij}-0.5$ for $p^{e,ij}=1,2,3,....,100$. With the payoff function given in (3), the expected payoff is[29]

$$\mathbb{E}(u_t^i|p_t^{e,i}=p^{e,ij},Pr(p_t^e))= \\ \sum_{p^{ej}=1}^{100} Pr(p_t^e=p^{ej})\cdot(1-\frac{1}{49}((c+b\cdot(p^{ej}-0.5)-(p^{e,ij}-0.5))^2+0.25)) \tag{25}$$

where the last term comes from the variance of the exogenous disturbance $\epsilon_t \sim N(0,0.25)$. For every λ, a fixed point is defined for the probability distribution $Pr(p_t^{e,i}=p^{e,ij})$.[30] This fixed point is obtained through value function iteration. Since (25) contains the average action p^{ej} 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 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

[26]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 action corresponds to its density so that $Pr(a_t^i)=f(a_t^i)$. Since under the division of bins used here, QRE, the Bayesian logit and TS all create densities, their log-likelihoods are comparable.

[27]Less than 1 % of all price forecasts are greater than 100.

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

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

[30]For the treatments with the robot traders, p^{ej} 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.

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 λ 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 given in (20). This perception (or belief structure) is used for the calculation of the expected payoff for a particular forecast $p^{e,i}$, $V(.) = \mathbb{E}(u_t^i | N(\bar{p}_t^*, \rho_t), p_t^{e,i} = p^{e,i})$, in the logit expression, where the payoff is given by equation (18).

Proposition 2. *The expected payoff for a particular forecast $p^{e,i}$ conditional on the perception in* (20) *is given by*

$$\mathbb{E}_t(u_t^i | N(\bar{p}_t^*, \rho_t), p_t^{e,i} = p^{e,i}) = 1300 - \frac{1300}{49}[\sigma^2 + \rho_{t-1} + \bar{p}_t^{*^2} - 2\bar{p}_t^* p^{e,i} + p^{e,i^2}] \tag{26}$$

where $\bar{p}_{t-1}^*$ denotes the expectation of p^* and ρ_t the posterior variance using information up to period t-1.[31]

Proof. See Appendix 6.2. □

Equation (26) reveals why the Bayesian logit predicts no heteroskedasticity in behavior. The probability of choosing a price prediction $p^{e,i}$ over any alternative price prediction $\hat{p}^{e,i}$ is determined by the cardinal differences in utility $u(p^{e,i}) - u(\hat{p}^{e,i})$. It is easy to see that ρ_{t-1}, the only possible source of volatility differences, cancels out. The only dynamic effect is through the agent's estimate $\bar{p}_t^*$, which affects the mean of the agent's choice over time but not the variance.

Analogously to Thompson Sampling, $\bar{p}_t^*$ is updated using the Kalman filter as in equation (21) and ρ_t is updated as in equation (23). *Since the Kalman filter is an adaptive filter, the Bayesian logit corresponds to a particular case of adaptive learning with exogenous shocks.* The expected

[31] The max(.) is ignored here for analytical tractability. Only few forecasts (approx. 6.2 %) yielded a payoff of zero.

payoff is inserted into the logit equation. The integral of $\exp(\lambda \mathbb{E}_t(u_t^i | N(\bar{p}_t^*, \rho_t), p_t^{e,i} = p^{e,i}))$ has been evaluated by a discrete approximation, dividing the action space from 0 to 100 up into equal bins.[32] Altogether, three parameters need to be estimated for the Bayesian logit: the two initial priors from Thompson Sampling, $\bar{p}_0^*, \rho_0$, the perceived noise variance σ^2, as well as the rationality parameter λ from the logit distribution.[33]

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. Strictly spoken, rational expectations can therefore be ipso facto rejected, although the results by Heemeijer et al. (2009) clearly indicate that rational expectations can be considered a good long-run predictor under strategic substitutes.

In terms of in-sample comparison, there is a clear ranking between the models. Thompson Sampling provides the best fit; the Bayesian logit provides the second-best fit; QRE provides the third-best fit and all three models beat a random uniform. This ranking is preserved under the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). All differences are significant both for the in-sample likelihoods and the BICs. The in-sample statistics and the results of the statistical tests are shown in appendix 6.1.2.

The ranking between the models that has been established in-sample is preserved for out-of-sample predictions, the toughest test. Figure 5 shows the resulting out-of-sample likelihoods and table 6 shows the results for the pairwise Wilcoxon signed-rank tests. The following observations stand out:

Observation 3. All three models perform significantly better than a random draw on the interval [0,100].

Observation 4. Thompson Sampling provides the best out-of-sample fit.

Observation 5. Both Thompson Sampling and the Bayesian logit provide a better out-of-sample fit than QRE.

[32] The midpoint of every interval has been used for expected payoff calculation.

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

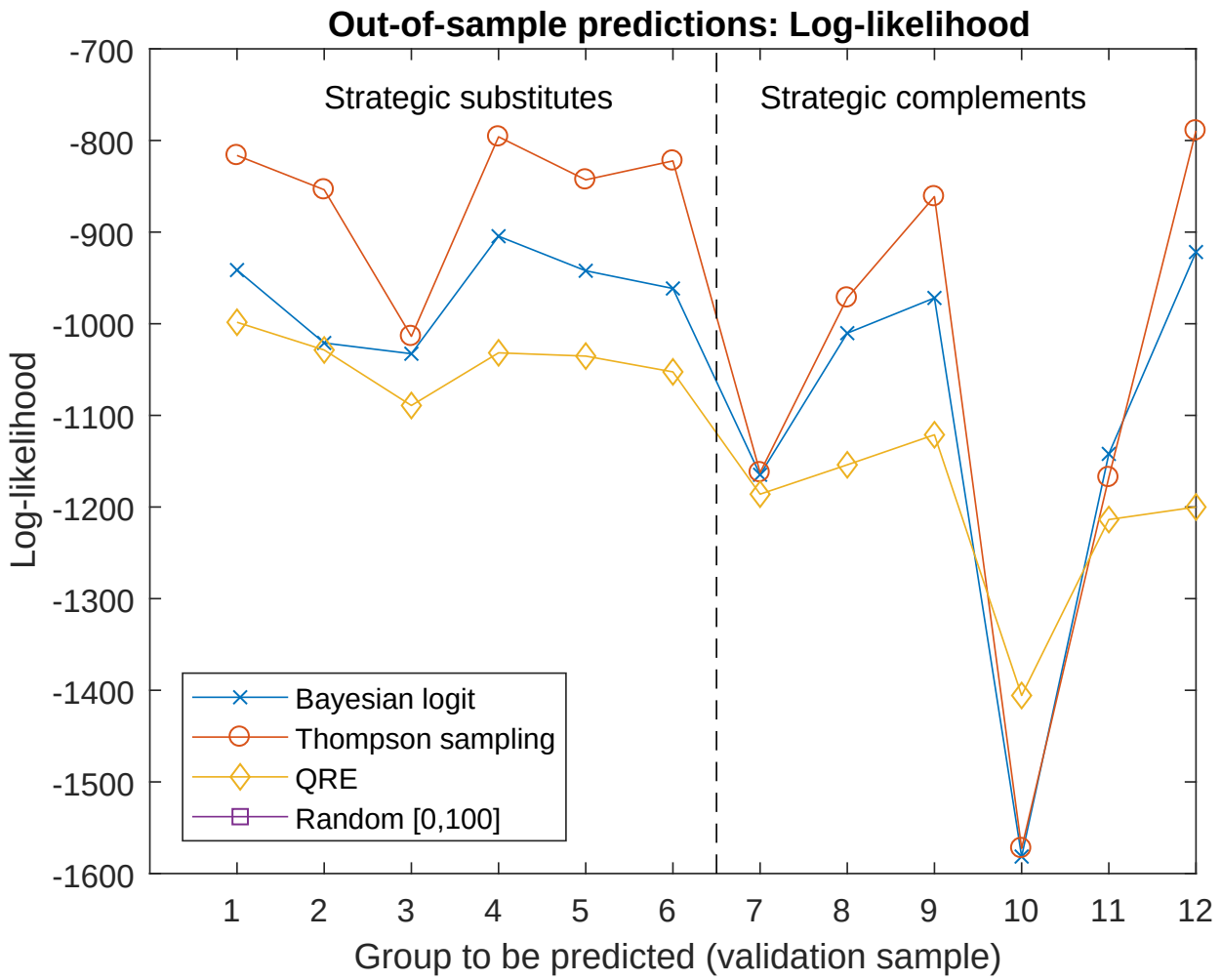

Figure 5: Learning to forecast: Likelihoods of validation samples

The fact that Thompson Sampling provides better out-of-sample predictions in 11 out of 12 groups renders it the preferred model by the signed-rank test.

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.

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0121)	-	-	-
QRE	TS (0.0121)	Logit (0.0278)	-	-
Random	TS (0.0029)	Logit (0.0048)	QRE (0.0029)	-

Table 6: Learning to forecast: Preferred model for out-of-sample forecasting by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

Parameters

Observation 6. 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 λ 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.

	prior mean $\bar{p}_0^*$	prior variance ρ_0	noise variance σ^2	rationality λ
Logit	43.03	148.23	26.26	0.48
	(0.88)	(0.47)	(3.64)	(0.05)
TS	62.38	145.12	1,171.14	-
	(0.39)	(10.29)	(133.99)	
QRE	-	-	-	0.22
				(0.06)

standard errors in parentheses

Table 7: Learning-to-forecast experiment: MLE parameter estimates

4.5 Endogenous noise variance

As shown by equation 23, the series of posterior variances is deterministically pinned down. As Thompson Sampling introduces randomness through sampling from the posterior, there are by design no differences in randomness across environments - only over time. 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 σ^2.

Conceptually, the case of both unknown mean and variance is not different from the case with only an unknown mean. For the sake of tractability, a conjugate prior is used for a Gaussian distribution with unknown mean and variance. In the exposition of this conjugate prior, I closely follow Hoff (2010). In the previous section, the conjugate prior for the mean p^* was a normal of the form $N(\bar{p}_0^*, \rho_0)$. One can consider a prior variance of $\rho_0 = \sigma^2/\kappa_0$, which gives $\bar{p}_0^*$ and κ_0 the interpretation of the mean and sample size of a prior sample that agents have in mind at the beginning. For σ^2, a conjugate prior is needed that is defined on the interval $(0, \infty)$. It turns out that the conjugate prior for the precision, $\frac{1}{\sigma^2}$, is the gamma-distribution, so that the conjugate prior for an unknown variance is the inverse gamma distribution. (See e.g. Gelman et al. (2013).) One can adopt the following parameterization

$$\sigma^2 \sim \text{inverse-gamma}(\frac{\upsilon_0}{2}, \frac{\upsilon_0}{2}\sigma_0^2) \tag{27}$$

υ_0 corresponds to the size of a prior (hypothetical) sample that agents have in mind before playing and σ_0^2 corresponds to the sample variance of this prior sample. Distributions of this family have an expected value being $E(\sigma^2) = \frac{\upsilon_0/2}{\upsilon_0/2-1}\sigma_0^2$ for this parameterization.

Since it is implausible that agents use fewer observations to estimate the variance than the mean (or vice versa) when a hypothetical, prior sample of a specific size is available, I impose the restriction that $\kappa_0 = \upsilon_0$. This assumption is frequently made in practical applications of Bayesian learning with unknown mean and variance of a Gaussian distribution and is referred to as Hoff's conjugate prior. (Hoff, 2010) This also has the advantage that the number of exogenous parameters is reduced to three: $\bar{p}_0^*$, υ_0 and σ_0^2. Bayes' rule implies that these parameters are updated after

period t the following way (see Hoff (2010) for derivations):

$$\bar{p}_t^* = \frac{\upsilon_0 \bar{p}_0^* + \sum_{s=1}^{t} p_s}{\upsilon_0 + t} \tag{28}$$

$$\upsilon_t = \upsilon_0 + t \tag{29}$$

$$\sigma_t^2 = \frac{1}{\upsilon_t}[\upsilon_0 \sigma_0^2 + (t-1)\delta_t^2 + \frac{\upsilon_0 t}{\upsilon_t}(\frac{1}{t}\sum_{s=1}^{t} p_s - \bar{p}_0^*)^2] \tag{30}$$

where $\delta_t^2 = \frac{\sum_{s=1}^{t}(p_s - \frac{1}{t}\sum_{w=1}^{t} p_w)^2}{t-1}$ is the sample variance. It is easy to see that (28) is the sample mean of prior and actual observations and (29) is the cumulative sample size of actual and prior observations. In (30), the terms in the square brackets capture the "prior sums of squared deviations plus the data sum of squared deviations." The term $\upsilon_0 \sigma_0^2$ corresponds to the prior sums of squares and the term $(t-1)\delta_t^2$ to the data sums of squares. The third term is slightly more tedious to interpret: it is an estimate of the variance using the prior mean $\bar{p}_0^*$ and the observations from the data, being information that ought to also be used for the posterior.

The Bayesian logit adds trembles in the action spaces but otherwise makes optimal use to form the beliefs. Thus, the estimate for the mean in period t is given by $E_t(p^*) = \bar{p}_t^*$, the estimate for the variance of p_t by $E_t(\sigma^2) = \sigma_t^2 \frac{\upsilon_t/2}{\upsilon_t/2-1}$ and the estimate for the posterior variance of p^* by $\rho_t = \frac{E_t(\sigma^2)}{\upsilon_t}$. These derivations can be inserted into equation (26), which is then used to calculate the likelihood for the Bayesian logit.

Under Thompson Sampling, agents draw a value for the variance from the inverse-gamma distribution, conditional on which they draw a value for the mean from the normal distribution. From the observed data, there is no possibility to directly infer agents' estimate for the unknown variance. Thus, in order to apply Thompson Sampling the unknown variance needs to be marginalized out to yield a marginal posterior. A marginal posterior is defined as $g(p^*|\bar{p}_t^*, \upsilon_t) = \int h(p^*, \sigma^2|\bar{p}_t^*, \upsilon_t)\partial\sigma^2$. Drawing an estimate for the mean from the marginal posterior is isomorphic to drawing the variance from the inverse-gamma distribution and drawing the mean from the normal. Hoff (2010) shows that the marginal posterior distribution of p^* at time t can be obtained in a closed form: $\frac{p^* - \bar{p}_t^*}{\sigma_t/\sqrt{\upsilon_t}}$ follows a t-distribution with υ_t degrees of freedom. We apply the following theorem (see e.g. Hoff (2010), p.231):

Theorem 7. *If a random variable X has density function $f(x)$ and $y = r(x)$ is a continuous transformation that is either strictly increasing or strictly decreasing, then $Y = r(x)$ has density*

function

$$g(y) = f(s(y))|\frac{\partial s(y)}{\partial y}| \tag{31}$$

where $s(y) = r^{-1}(y)$ *is the inverse function of* $r(x)$.

Denoting $X_t = \frac{p^* - \bar{p}_t^*}{\sigma_t/\sqrt{\upsilon_t}}$ so that $p^* = \sigma/\sqrt{\upsilon_t} \cdot X_t + \bar{p}_t^*$, one can establish by using the above theorem that the marginal posterior of p^* is a non-standard student distribution:

$$g(p^*|\bar{p}_t^*, \upsilon_t, s_t) = \frac{\Gamma((\upsilon_t + 1)/2)}{\sqrt{\pi\upsilon_t}\Gamma(\upsilon_t/2)s_t}(1 + \frac{((p^* - \bar{p}_t^*)/s_t)^2}{\upsilon_t})^{-(\upsilon_t+1)/2} \tag{32}$$

where $\Gamma(.)$ is Euler's Gamma function and $s_t \equiv \sigma_t/\sqrt{\upsilon_t}$. To implement Thompson Sampling, a subjective estimate $\tilde{p}_t^{*,i}$ is drawn from this marginal posterior.

Methodology Just as in section 4.3, three parameters need to be estimated for TS ($\bar{p}_0^*$, υ_0, σ_0^2) and four for the Bayesian logit ($\bar{p}_0^*$, υ_0, σ_0^2, λ). Looking at equation (26) and recalling that the probability of choice in a logit model is determined by the utility differences, it becomes apparent that an endogenous noise variance σ_t^2 does not explain variance in behavior over time and across setups. The only restriction that has been made for the estimation is that $\upsilon_0 \geq 2$, since a sample of only one observation would automatically yield a variance estimate of zero.

4.5.1 Estimation results

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0060)	-	-	-
QRE	TS (0.0060)	Logit (0.0278)	-	-
Random	TS (0.0029)	Logit (0.0029)	QRE (0.0029)	-

Table 8: Learning to forecast with endogenous variance: Preferred model for out-of-sample forecasting by the Wilcoxon signed-rank tests in pairwise comparison (p-values in parentheses)

Model comparison Appendix 6.1.3 shows the in-sample statistics (likelihoods, AICS and BICs) as well as the results of the test comparing the measures. Similar to the case with the Kalman filter, there is a clear ranking. Thompson Sampling provides the best in-sample fit, no matter whether measured by the in-sample likelihood or whether a penalty criterion like AIC or BIC is

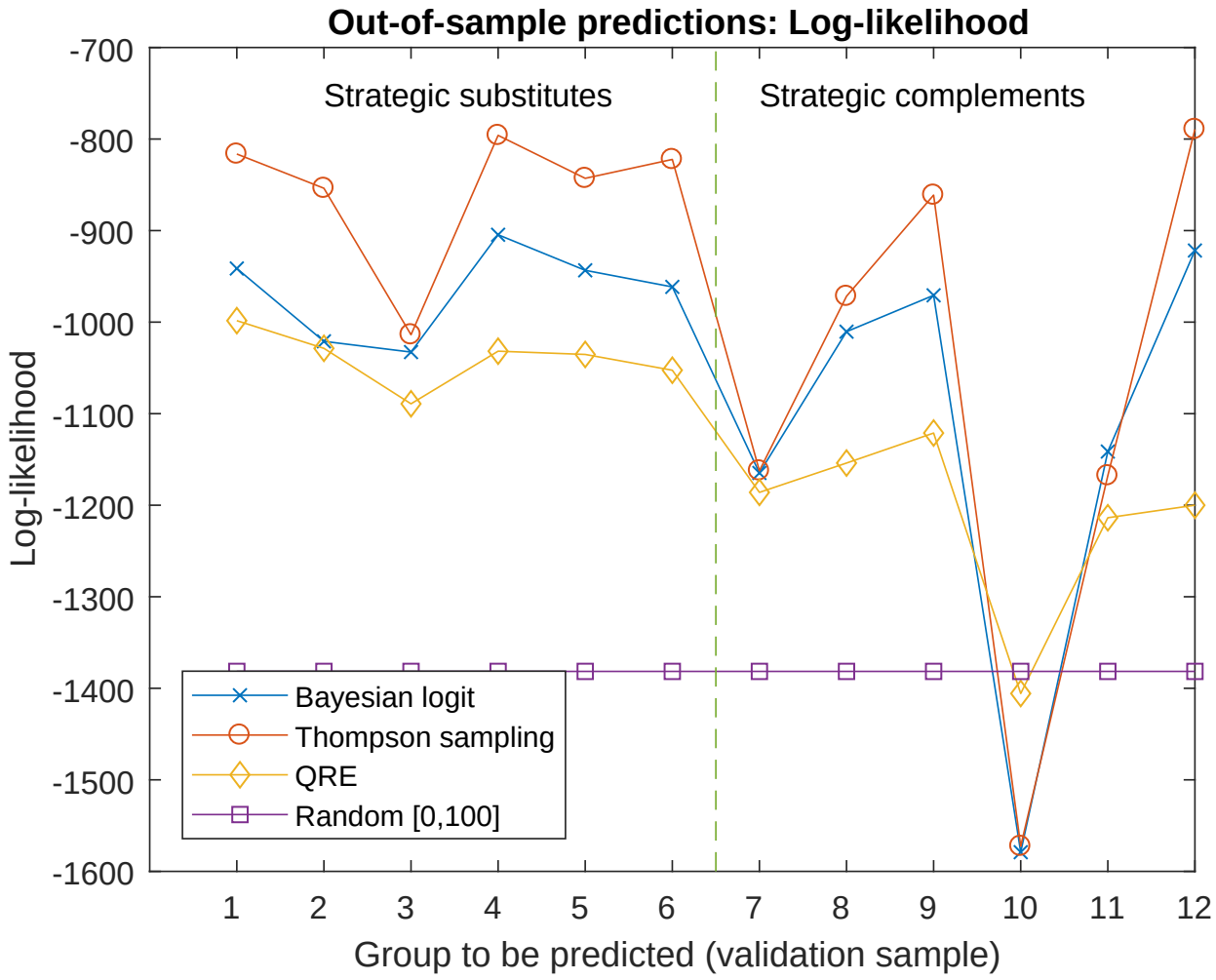

Figure 6: Learning to forecast with endogenous variance: Likelihoods of validation samples

used. The second best fit is given by the Bayesian logit, which provides a significantly better fit than QRE. All three models provide a significantly better fit than a random uniform.

A Wilcoxon signed-rank test over the experimental groups reveals that the empirical fit of TS with endogenous noise variance is better than the standard case with constant noise variance (p-value: 0.0188). Figure 6 shows the out-of-sample fits. Thompson Sampling provides a better fit than both QRE and the Bayesian logit in all groups except for group 11. Table 6 shows the results of pairwise Wilcoxon signed-rank tests.

Observation 8. The empirical out-of-sample fit of TS is significantly better than the Bayesian logit and QRE.

Parameters As shown in table 9, the estimates for the rationality parameter λ are similar to the standard Kalman filtering case. Likewise, the value for the prior mean for Thompson Sampling is robust to incorporating Bayesian learning about the variance.

	prior mean $\bar{p}_0^*$	prior sample variance σ_0^2	rationality λ	prior sample size υ_0
Logit	56.51	903.27	0.45	21.04
	(2.79)	(10.72)	(0.05)	(20.80)
TS	63.16	2,220.10	-	14.14
	(0.37)	(282.44)		(0.84)
QRE			0.22	-
			(0.06)	

standard errors in parentheses

Table 9: Learning-to-forecast with endogenous variance: MLE parameter estimates

The estimates for the variance in the endogenous variance case are higher than in the constant variance case. This is because an endogenous variance is flexible enough to allow, for example, for a high variance at the beginning, which declines as time passes. A (misspecified) constant variance not only has to account for observations with high variance at the beginning but also observations with low variance later on. Hence, this explains why the estimate of a variance that is forced to be constant is lower than the initial value of endogenous variance.

5 Discussion

This paper has introduced Thompson Sampling, a learning 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 number of independent observations. Hence, it would be intriguing to test Thompson Sampling for different datasets that may contain observational data.

Third, a specific theory about agents' priors has been applied to 2x2 games in this paper, making use of the equilibria and level-1 play for initialization. While this theory is backed by the data, there is certainly potential for more research on modeling and explaining agents' prior belief formation.

Fourth, Thompson Sampling can potentially have numerous policy implications. The fact that Thompson Sampling provides an empirically valid and tractable description of individual beliefs 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

Klaus Adam and Albert A. Marcet. Internal rationality, imperfect market knowledge and asset prices. *Journal of Economic Theory*, 146(3):1224–1252, 2011.

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *JMLR Workshop and Conference Proceedings (COLT2012)*, 23:39.1–39.26, 2012a.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 28:127–135, sep 2012b.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, dec 1974.

Larbi Alaoui and Antonio Penta. Endogenous Depth of Reasoning. *Review of Economic Studies*, 83(4):1297–1333, 2016.

Mikhail Anufriev and Cars Hommes. Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments. *American Economic Journal: Microeconomics*, 4 (225408):35–64, 2012.

Jose Apesteguia and Miguel A Ballester. Monotone Stochastic Choice Models: The Case of Risk and Time Preferences. *Journal of Political Economy*, 126(1):695504, 2017.

Jasmina Arifovic and John Ledyard. Scaling Up Learning in Public Good Games. *Journal of Public Economics*, 6(2):203–238, 2004.

Timothy EJ Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–21, 2007.

Daniel Bernoulli. Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22(1): 23, jan 1954.

H. D. Block and Jacob Marshak. Random Orderings and Stochastic Theories of Response. In Ingram Olkin, Sudhist G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, editors, *Contributions to probability and statistics*. Stanford University Press, Stanford, CA, 1960.

Yves Breitmoser. Strategic reasoning in p-beauty contests. *Games and Economic Behavior*, 75 (2):555–569, 2012.

Monica Brezzi and Tze Leung Lai. Incomplete Learning from Endogenous Data in Dynamic Allocation. *Econometrica*, 68(6):1511–1516, nov 2000.

William A Brock and Cars H Hommes. A Rational Route to Randomness. *Econometrica*, 65(5): 1059–1095, 1997.

George W. Brown. Iterative Solutions of Games by Fictitious Play. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 374–376. Wiley, New York, 1951.

J R Busemeyer and Y M Wang. Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *J Math Psychol*, 44(1):171–189, 2000.

C. F. Camerer and Ernst Fehr. When Does "Economic Man" Dominate Social Behavior? *Science*, 311(5757):47–52, 2006.

Colin Camerer and Teck-hua Ho. Experienced-Weighted Attraction Learning in Normal Form Games. *Econometrica*, 67(4):827–874, 1999.

Colin F Camerer, Teck-hua Ho, and Juin-Kuan Chong. A Cognitive Hierarchy Model. *Quarterly Journal of Economics*, 119(3):861–898, 2004.

A. Colin Cameron and Pravin Triverdi. *Microeconometrics: Methods and Applications*. Cambridge University Press, New York, 2005.

Olivier Chapelle and Lihong Li. An Empirical Evaluation of Thompson Sampling. *Advances in Neural Information Processing Systems*, pages 2249—-2257, 2011.

Nick Chater and Christopher D. Manning. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344, 2006.

Chris Chatfield. Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15(7):495–508, 1996.

Yin-wong Cheung and Daniel Friedman. Individual Learning in Normal Form Games :. *Games and Economic Behavior*, 19:46–76, 1997.

Todd Clark and Michael McCracken. *Advances in forecast evaluation*, volume 2. Elsevier B.V., 2013. ISBN 9780444627315.

Miguel A. Costa-Gomes and Georg Weizsäcker. Stated Beliefs and Play in Normal-Form Games. *Review of Economic Studies*, 75(3):729–762, jul 2008.

James Costain and Anton Nakov. Logit price dynamics. *CEPR Discussion Paper No. 10731*, 2015.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

Stefano DellaVigna. Structural Behavioral Economics. In Doug Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics - Foundations and Applications 1*, pages 613–723. Elsevier, 1st edition, 2018.

Martin Dufwenberg, Uri Gneezy, Jacob K. Goeree, and Rosemarie Nagel. Price floors and competition. *Economic Theory*, 33(1):211–224, 2007.

Ido Erev, Alvin E. Roth, Robert L. Slonim, and Greg Barron. Predictive value and the usefulness of game theoretic models. *International Journal of Forecasting*, 18(3):359–368, 2002.

Ido Erev, Alvin E. Roth, Robert L. Slonim, and Greg Barron. Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory*, 33(1):29–51, 2007.

George W. Evans and Seppo Honkapohja. *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, jan 2001. ISBN 9781400824267.

George W Evans and Garey Ramey. Expectation Calculation and Macroeconomic Dynamics. *American Economic Review*, 82(1):207–224, 1992.

George W Evans, Seppo Honkapohja, and Noah Williams. Generalized stochastic gradient learning. *International Economic Review*, 51(1):237–262, 2010.

Denzil G. Fiebig, Michael P. Keane, Jordan Louviere, and Nada Wasi. The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science*, 29 (3):393–421, 2010.

Mira Frick, Ryota Iijima, and Tomasz Strzalecki. Dynamic Random Utility. *Cowles Foundation Discussion Paper No. 2092*, 2017.

John Frisby and James Stone. *Seeing: The Computational Approach to Biological Vision*. MIT Press, Cambridge and London, 2nd edition, 2013.

Drew Fudenberg and David Kreps. Learning Mixed Equilibria. *Games and Economic Behavior*, 5(3):320–367, 1993.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

Drew Fudenberg and Annie Liang. Predicting and Understanding Initial Play. *Working Paper*, 2017.

Andrew Gelman, John Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman Hall/CRC, Boca Raton, FL, 3rd edition, 2013.

Samuel J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

I. Gilboa and D. Schmeidler. Case-Based Decision Theory. *The Quarterly Journal of Economics*, 110(3):605–639, aug 1995.

Jeff Gill. *Generalized Linear Models: A Unified Approach*. SAGE Publications, Thousand Oaks, CA, 2001.

John Charles Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

J. K. Goeree and C. A. Holt. Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567, 1999.

Jacob K. Goeree and Charles A. Holt. An experimental study of costly coordination. *Games and Economic Behavior*, 51(2 SPEC. ISS.):349–364, 2005.

Jacob K. Goeree, Thomas R. Palfrey, Brian W. Rogers, and Richard D. McKelvey. Self-correcting information cascades. *Review of Economic Studies*, 74(3):733–762, 2007.

Jacob K. Goeree, Charles A. Holt, and Thomas R. Palfrey. *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press, Princeton, NJ, 2016.

William H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, New Jersey, 5th edition, 2002.

Faruk Gul and Wolfgang Pesendorfer. Random Expected Utility. *Econometrica*, 74(1):121–146, jan 2006.

Philip A Haile, A Hortaçsu, and G Kosenok. On the Empirical Content of Quantal Response EquilibriumVol. 98, No. 1 (Mar., 2008), pp. 180-200. *The American Economic Review*, 98(1): 180–200, 2008.

Peter R. Hansen. A winner's curse for econometric models: on the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection. *Manuscript, Department of Economics, Stanford ...*, 2010.

Peter Heemeijer, Cars Hommes, Joep Sonnemans, and Jan Tuinstra. Price stability and volatility in markets with positive and negative expectations feedback: An experimental investigation. *Journal of Economic Dynamics and Control*, 33:1052–1072, 2009.

Peter Hoff. *A First Course in Bayesian Statistical Methods*. Springer, New York, 1st edition, 2010.

Cars Hommes. *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems*. Cambridge University Press, Cambridge, 2013.

Ed Hopkins. Two Competing Models of How People Learn in Games. *Econometrica*, 70(6): 2141–2166, nov 2002.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35, 1960.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7568 LNAI, pages 199–213. 2012. ISBN 9783642341052.

T. Keasar, Ella Rashkovich, Dan Cohen, and Avi Shmida. Bees in two-armed bandit situations: foraging choices and possible decision mechanisms. *Behavioral Ecology*, 13(6):757–765, 2002.

Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th international joint conference on Artificial intelligence*, 2: 1137–1143, 1995.

Gary Koop and Dale J. Poirier. Bayesian analysis of logit models using natural conjugate priors. *Journal of Econometrics*, 56(3):323–340, 1993.

Edward E. Leamer. *Specification searches: Ad hoc inference with nonexperimental data*. Wiley, New York, 1978.

Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson Sampling is Asymptotically Optimal in General Environments. *Working paper*, 2018.

J. Scott Long. *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications, Thousand Oaks, CA, 1997.

Jordan Louviere and Thomas Eagle. Confound It! That Pesky Little Scale Constant Messes Up Our Convenient Assumptions. *Sawtooth Software Conference*, (September):211–228, 2006.

Ramon Marimon and Shyam Sunder. Expectations and learning under alternative monetary regimes: an experimental approach. *Economic Theory*, 4(1):131–162, jan 1994.

Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012.

Daniel McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behaviour*, 10:6–38, 1995.

Othon M. Moreno and Yaroslav Rosokha. Learning under compound risk vs. learning under ambiguity – an experiment. *Journal of Risk and Uncertainty*, 53(2-3):137–162, 2016.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326, 1995.

Rosemarie Nagel and Nicolaas J. Vriend. An experimental study of adaptive behavior in an oligopolistic market game. *Journal of Evolutionary Economics*, 9(1):27–65, 1999.

Rosemarie Nagel, Christoph Bühren, and Björn Frank. Inspired and inspiring: Hervé Moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*, pages 1–17, 2016.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

John P O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304:452–54, 2004.

Martin J Osborne. *Introduction to Game Theory*. Oxford University Press, Oxford, 1st edition, 2003.

Elise Payzan-LeNestour and Peter Bossaerts. Learning about Unstable, Publicly Unobservable Payoffs. *Review of Financial Studies*, 28(7):1874–1913, 2015.

Alvin Roth and Ido Erev. Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review*, 88(4):848–881, 1998.

Alvin E. Roth and Michael W. Malouf. Game-theoretic models and the role of information in bargaining. *Psychological Review*, 86(6):574–594, 1979.

Adam N. Sanborn and Nick Chater. Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, 20(12):883–893, 2016.

Leonard J. Savage. *The Foundations of Statistics*. John Wiley & Sons, New York, 1954.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

Steven L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, nov 2010.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2):351–367, 2015.

Dale O Stahl. Evolution of Smartn Players. *Games and Economic Behavior*, 5(4):604–617, oct 1993.

James Stock and Mark W Watson. *Introduction to econometrics*. Pearson Education Limited, Essex, 3rd edition, 2015. ISBN 0-02-374545-2.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285, 1933.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge University Press, New York, 2nd edition, 2009. ISBN 9780521747387.

Bernhard Treutwein. Adaptive psychophysical procedures. *Vision Research*, 35(17):2503–2522, sep 1995.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

Nir Vulkan. An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, 14(1):101–118, 2000.

Joan Walker and Moshe Ben-Akiva. Generalized random utility model. *Mathematical social sciences*, 43(3):303–343, 2002.

Richard F. West and Keith E. Stanovich. Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory and Cognition*, 31(2):243–251, 2003.

Jon Williamson. Review of Bruno de Finetti's 'Philosophical Lectures on Probability'. *Philosophia Mathematica*, 18(1):130–135, 2010.

Daniel M. Wolpert. Probabilistic models in human sensorimotor control. *Human Movement Science*, 26(4):511–524, aug 2007.

Michael Woodford. Monetary Policy Analysis When Planning Horizons Are Finite. *Working Paper*, 2018.

David R. Wozny, Ulrik R. Beierholm, and Ladan Shams. Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, 6(8), 2010.

6 Appendix

6.1 In-sample fit

6.1.1 2x2 games

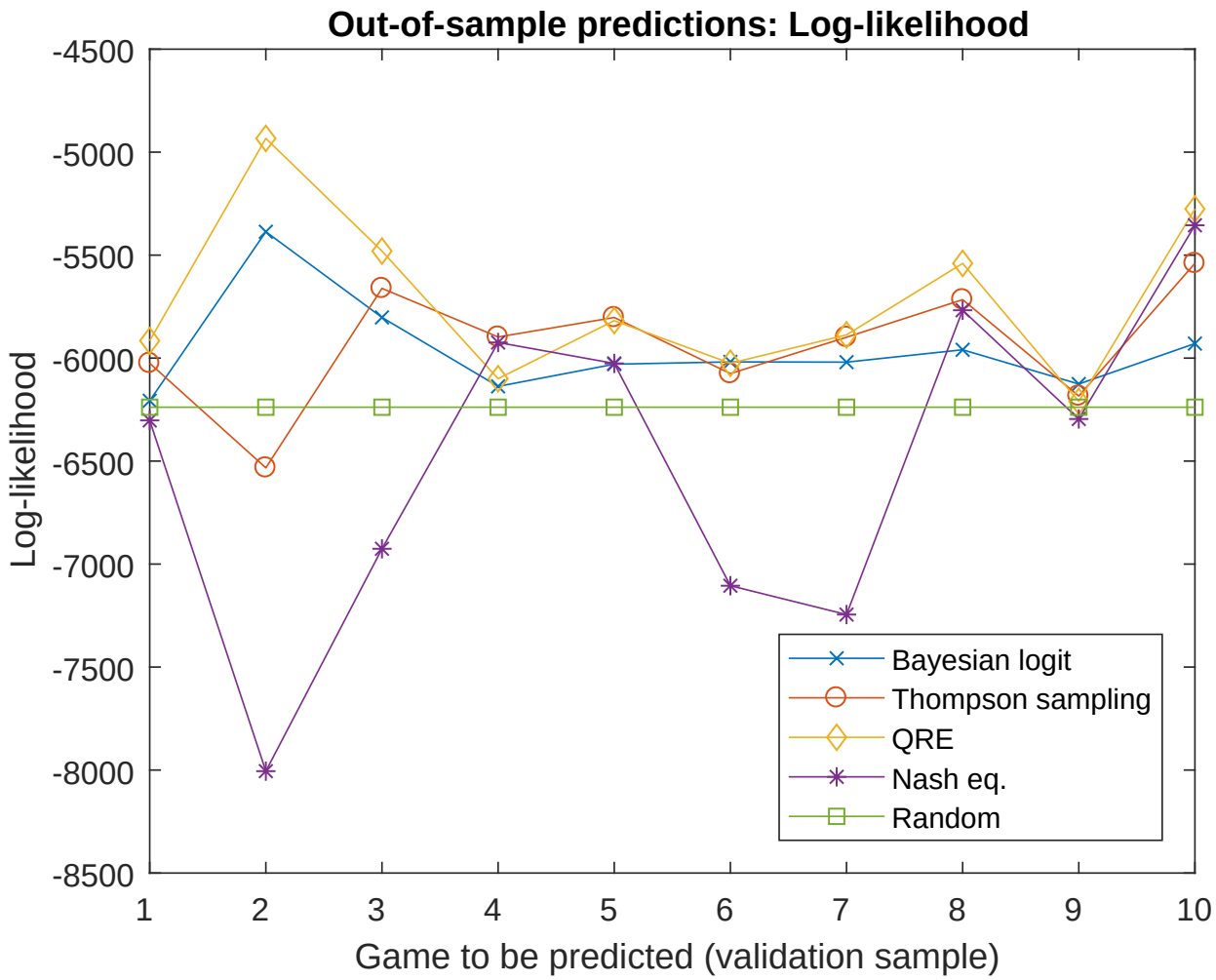

Figure 7: 2x2 games: Likelihoods of the in-sample fits (individual behavior)

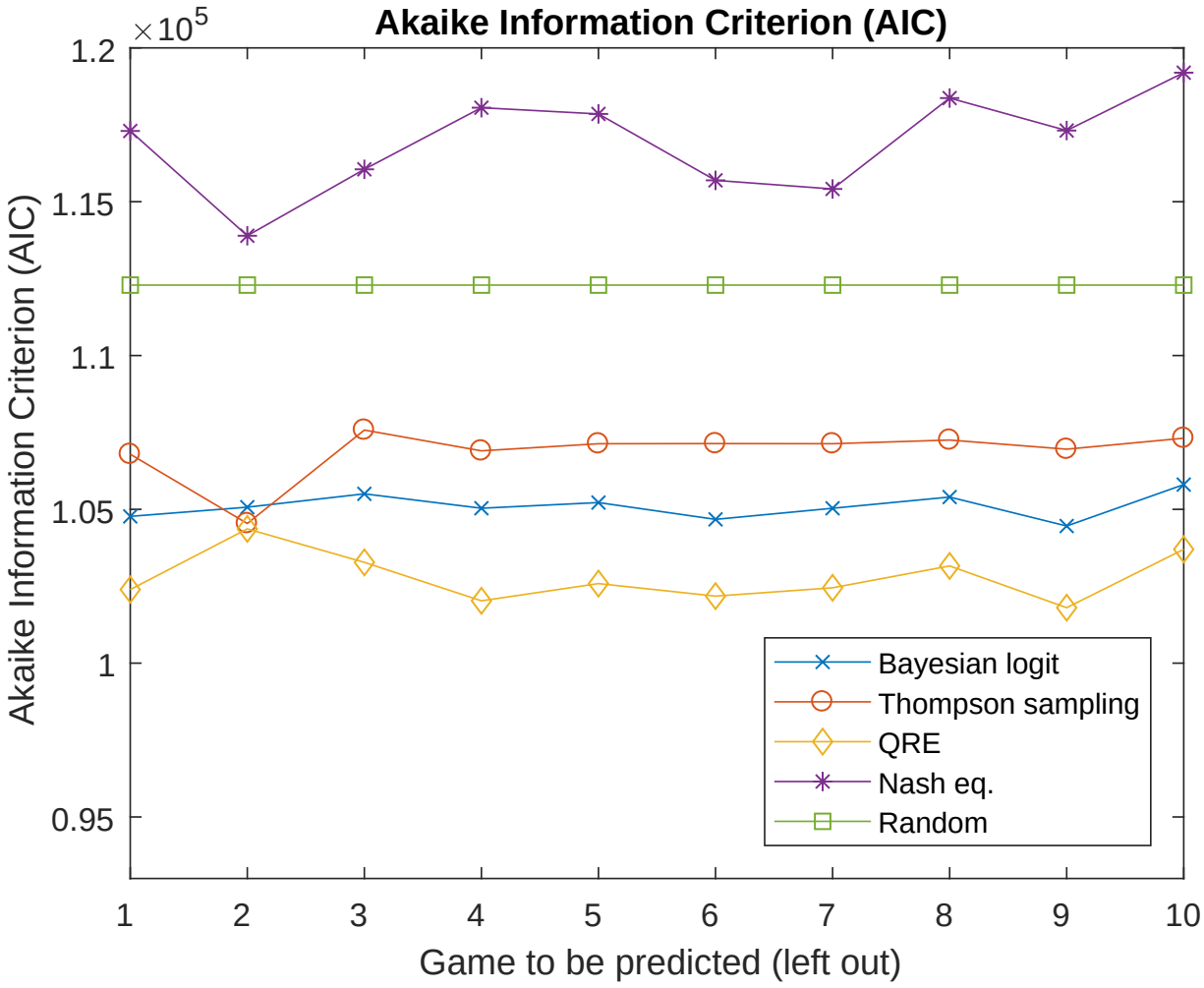

Figure 8: 2x2 games: Akaike Information Criteria (AICs)

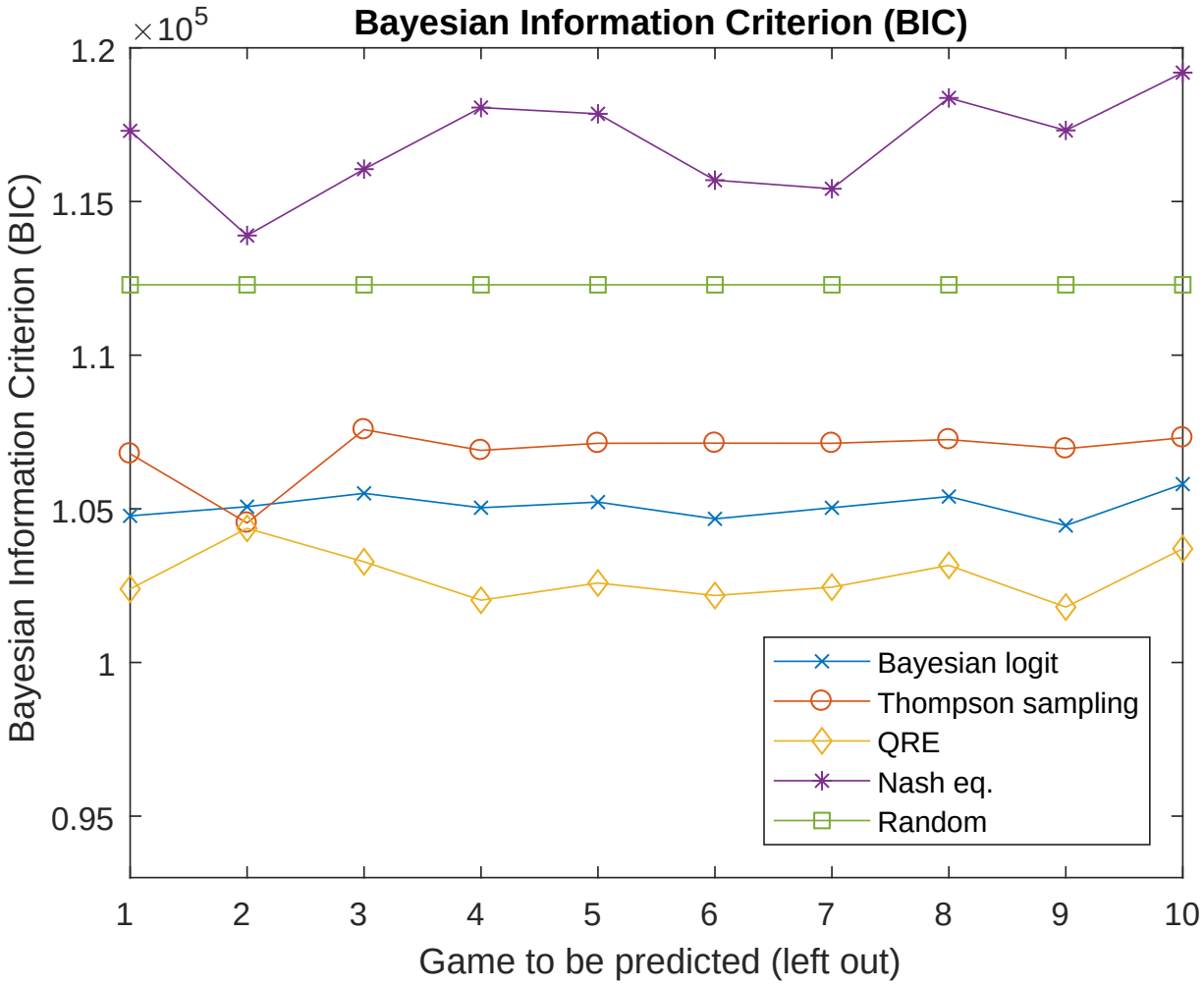

Figure 9: 2x2 games: Bayesian Information Criteria (BICs)

In-sample likelihood:

	TS	**Logit**	**QRE**	**Random**	**Nash**
TS	-	-	-	-	-
Logit	TS (0.0069)	-	-	-	-
QRE	QRE (0.0051)	QRE (0.0051)	-	-	-
Random	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	-	-
Nash	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	Random (0.0051)	-

Akaike Information Criterion (AIC):

	TS	**Logit**	**QRE**	**Random**	**Nash**
TS	-	-	-	-	-
Logit	TS (0.0069)	-	-	-	-
QRE	QRE (0.0051)	QRE (0.0051)	-	-	-
Random	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	-	-
Nash	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	Random (0.0051)	-

Bayesian Information Criterion (BIC):

	TS	**Logit**	**QRE**	**Random**	**Nash**
TS	-	-	-	-	-
Logit	TS (0.0069)	-	-	-	-
QRE	QRE (0.0051)	QRE (0.0051)	-	-	-
Random	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	-	-
Nash	TS (0.0051)	Logit (0.0051)	QRE (0.0051)	Random (0.0051)	-

Table 10: 2x2 games: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison for different measures of the in-sample fit (p-values in parentheses)

6.1.2 Expectation formation: Kalman filter

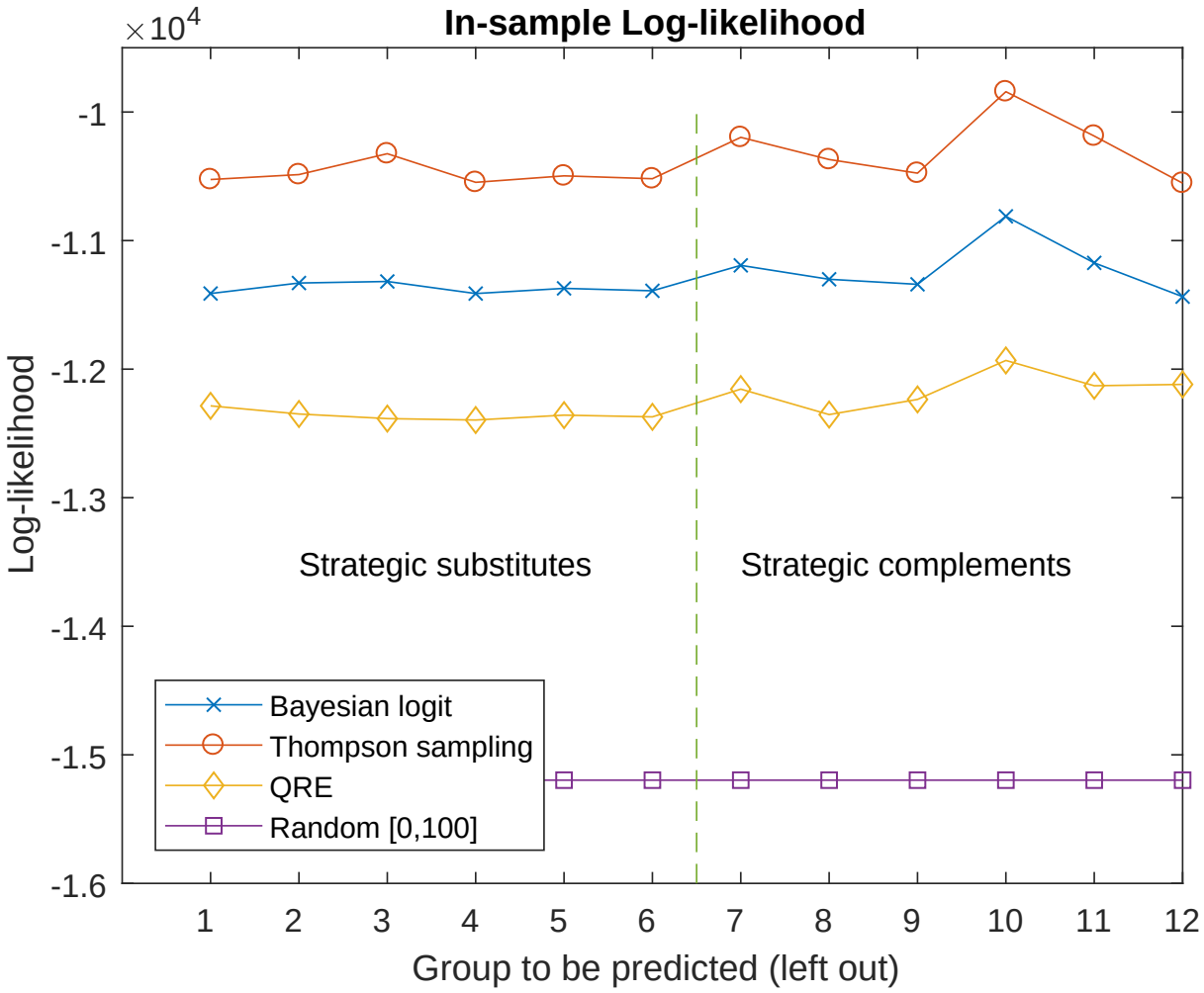

Figure 10: Learning to forecast: Likelihoods of in-sample fits

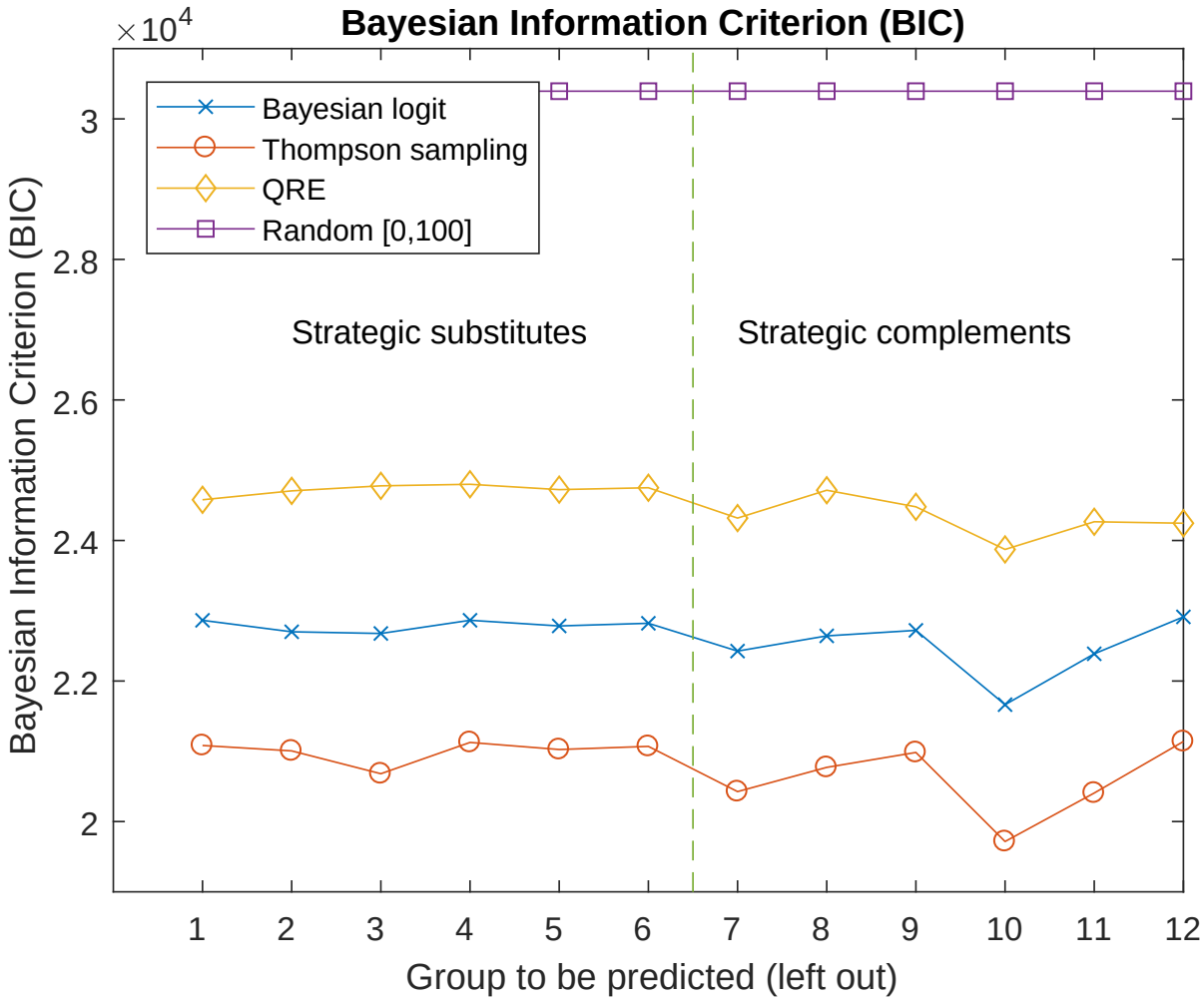

Figure 11: Learning to forecast: Bayesian Information Criterion (BIC) of in-sample fits

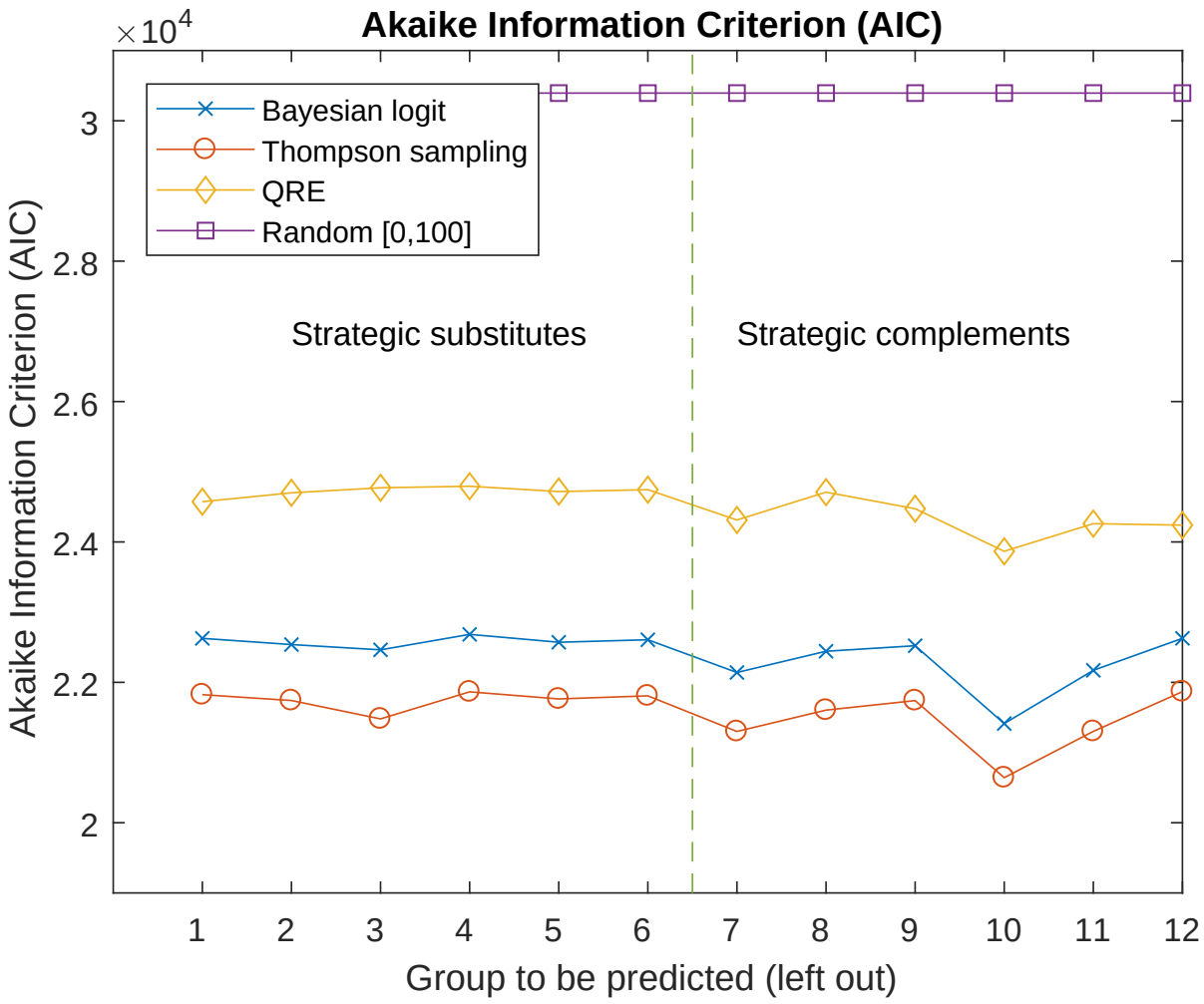

Figure 12: Learning to forecast: Akaike Information Criterion (AIC) of in-sample fits

In-sample Likelihood:

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Akaike Information Criterion (AIC):

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Bayesian Information Criterion (BIC):

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Table 11: Learning to forecast: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison for different measures of the in-sample fit (p-values in parentheses)

6.1.3 Expectation formation: Endogenous noise variance

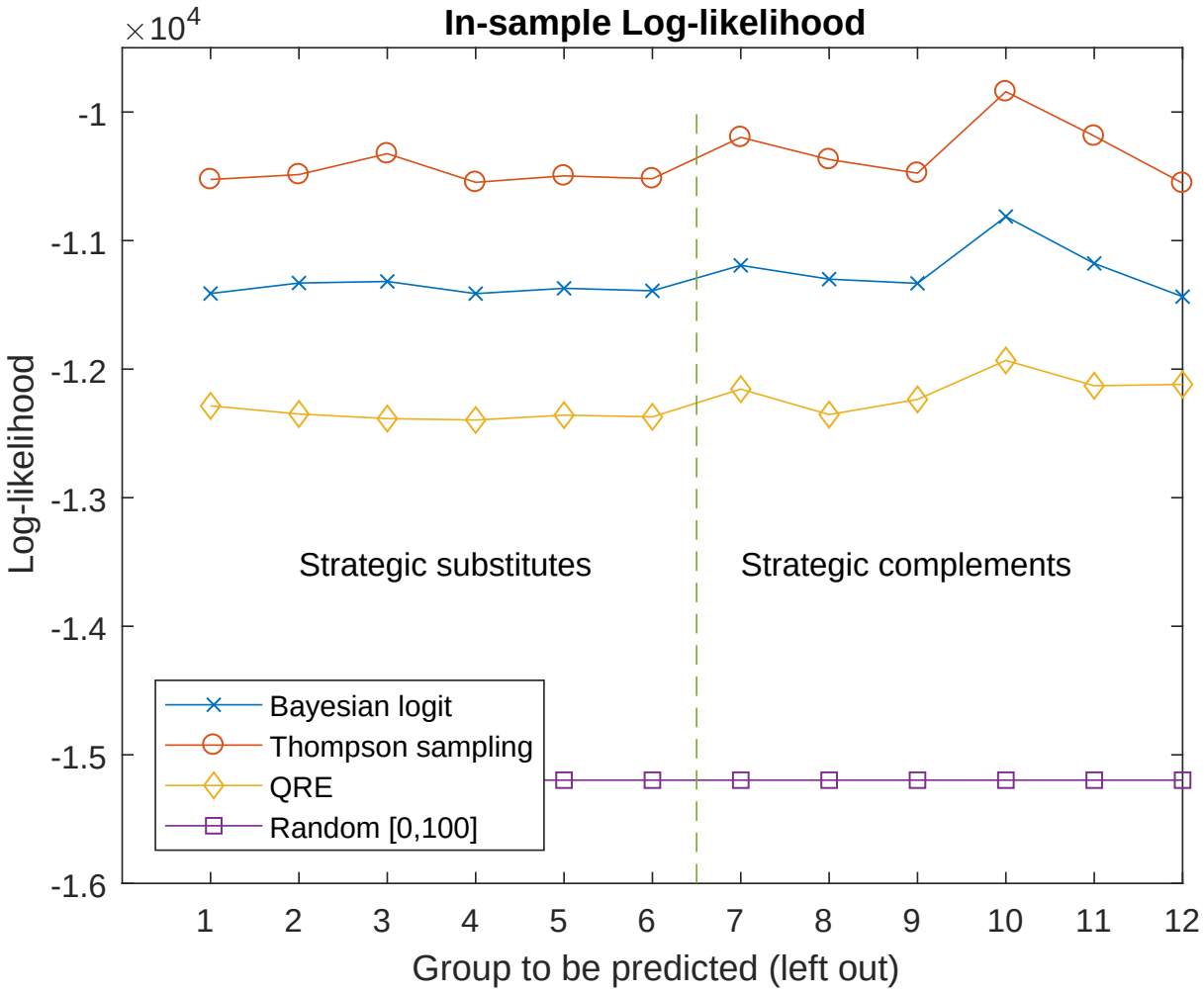

Figure 13: Learning to forecast: Likelihoods of in-sample fits with endogenous variance

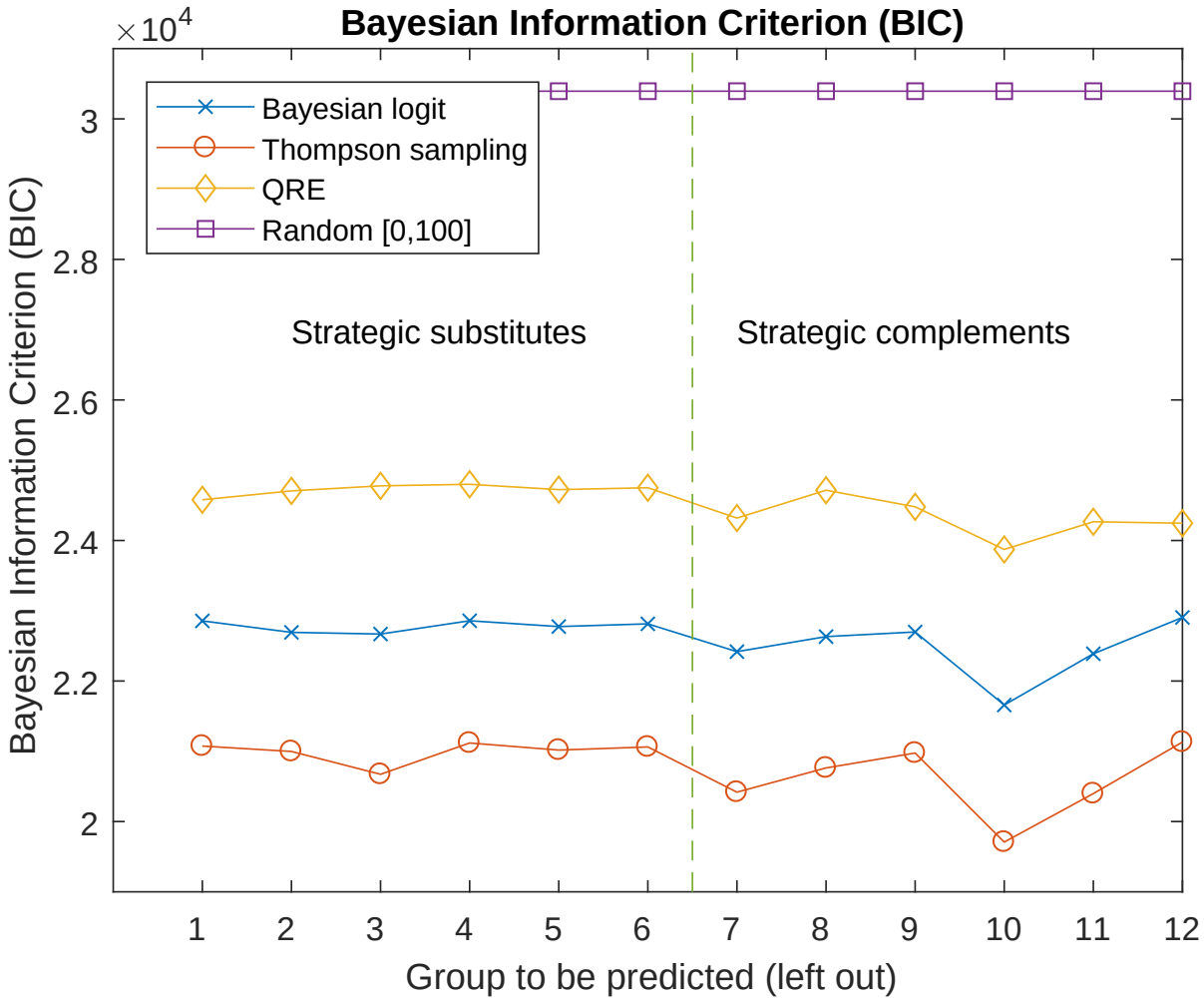

Figure 14: Learning to forecast: Bayesian Information Criterion (BIC) of in-sample fits with endogenous variance

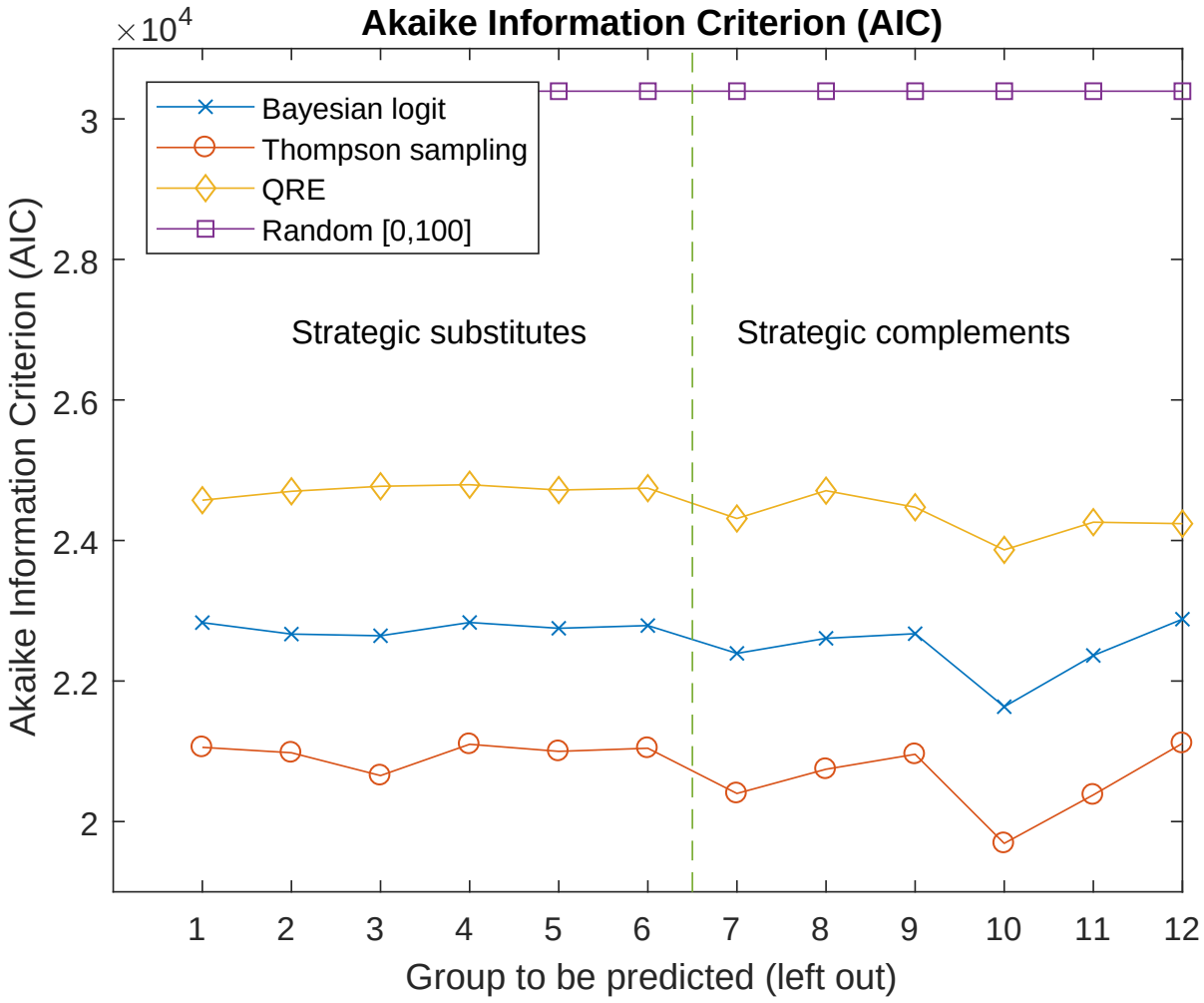

Figure 15: Learning to forecast: Akaike Information Criterion (AIC) of in-sample fits with endogenous variance

In-sample Likelihood:

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Akaike Information Criterion (AIC):

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Bayesian Information Criterion (BIC):

	TS	**Logit**	**QRE**	**Random**
TS	-	-	-	-
Logit	TS (0.0022)	-	-	-
QRE	TS (0.0022)	Logit (0.0022)	-	-
Random	TS (0.0022)	Logit (0.0022)	QRE (0.0022)	-

Table 12: Learning to forecast with endogenous variance: Preferred model by the Wilcoxon signed-rank tests in pairwise comparison for different measures of the in-sample fit (p-values in parentheses)

6.2 Proof of Proposition 2

Proof. Using (18), is:

$$u_t^i = 1300 - \frac{1300}{49}(p_t - p_t^{e,i})^2 \tag{33}$$

Under Gaussian perception given by (20), the score can be written as

$$u_t^i = 1300 - \frac{1300}{49}(p^* + \eta_t - p_t^{e,i})^2 \tag{34}$$

The expected value under this perception is

$$\mathbb{E}_t(u_t^i | N(\bar{p}_t^*, \rho_t), p_t^{e,i} = p^{e,i}) = 1300 - \frac{1300}{49} \underbrace{\mathbb{E}_t\{(p^* + \eta_t - p^{e,i})^2 | p_t^{e,i} = p^{e,i}\}}_{\equiv \xi_{t+1}} \tag{35}$$

Focusing only on the term ξ_{t+1}, we obtain

$$\begin{aligned}
\xi_{t+1} \equiv & \mathbb{E}_t\{(p^* + \eta_t - p^{e,i})^2 | p_t^{e,i} = p^{e,i}\} \\
= & \mathbb{E}_t\{(p^* - p^{e,i})^2\} + \mathbb{E}_t\{\eta_t^2\} \\
= & \mathbb{E}_t\{p^{*^2} - 2p^* \cdot p^{e,i} + p^{e,i^2}\} + \sigma^2 \\
= & \rho_{t-1} + \bar{p}_{t-1}^{*^2} - 2\bar{p}_{t-1}^* \cdot p^{e,i} + p^{e,i^2} + \sigma^2
\end{aligned} \tag{36}$$

where the last line uses the fact that $Var(\mu) = E(\mu^2) - (E(\mu))^2$. □

6.3 Augmenting Thompson Sampling for 2x2 games

Two ways of augmenting Thompson Sampling are considered: first, allowing decision-makers to weigh present observations differently from past observations; second, considering a more refined theory of the priors.

Generalization of Bayesian updating The unknown parameters are the priors at the beginning of the game α_0^{-i} and β_0^{-i}. Estimating those parameters under pure Bayesian belief updating yields high parameter estimates. A wide range of applications not only in the decision-making but also in the behavioral game theory literature explores the hypothesis of players having biased perceptions when updating. (Fudenberg and Levine, 1998; Roth and Erev, 1998; Camerer and

Ho, 1999) The approach taken here follows Goeree et al. (2007), who develop a generalization of Bayesian updating, and Moreno and Rosokha (2016), who generalize their framework to a setting with many time periods.

Bayes' rule as given by equation (1) can more generally be written as

$$D_t^i(\theta^i) = \frac{(\hat{\sigma}^{-i}(a_t^{-i}|\theta^i)))^{\xi^t} D_{t-1}^i(\theta^i)}{(\hat{\sigma}^{-i}(a_t^{-i}))^{\xi^t}} \tag{37}$$

The appeal of this specification is its flexibility given by the parameter ξ, which captures the perceived number of signals. Pure Bayesian learning is nested by setting $\xi = 1$. If the agent after observing the next signal acts as if she observed two signals, then $\xi = 2$. Values of $\xi > 0$ can be interpreted as limited memory, since agents pay more attention to more recent signals. Conversely, values of $\xi < 0$ can be interpreted as underweighting of the signal or "conservatism bias." To distinguish between old and more recent periods, following Moreno and Rosokha (2016), the weight of the signal in a period t is ξ^t, meaning that each new signal has ξ times the weight of the previous signal.

This implies an updating rule of

$$\alpha_t^{-i} = \begin{cases} \alpha_{t-1}^{-i} + \xi^t \cdot 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \alpha_t^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{38}$$

$$\beta_t^{-i} = \begin{cases} \beta_{t-1}^{-i} + \xi^t \cdot 1 & \text{if } a_t^{-i} = a^{-i,1} \\ \beta_{t-1}^{-i} & \text{if } a_t^{-i} = a^{-i,2} \end{cases} \tag{39}$$

6.3.1 Initial priors

In section 3.2, the learning rules have been treated to be the same across all games. Since there are 9 pairs playing every game, 9 estimations have been conducted in every game, leaving out one pair at the time. This procedure has the advantage that jackknife estimates for the standard errors are obtained as the sample standard deviation of the estimates.

The data display stark differences in initial play. It would thus be incorrect to use the initial conditions of one game to predict the dynamics of another game. However, if the variation in the priors is not understood, this will present an obstacle to forecasting the dynamics of new games ex-ante. Thus, the estimates of the prior parameters have been investigated for regularities.

While the estimates of these parameters for Thompson Sampling have a low variance, the estimates of the priors for the Bayesian logit have a high variance. For Thompson Sampling, the mean estimates $\hat{p}_0^{-i}(L) = \frac{\alpha_0^{-i}}{\beta_0^{-i}+\alpha_0^{-i}}$ have been found to be relatively close to the Nash equilibria of every game. Yet, one-sample Wilcoxon signed-rank tests find significant deviations from the Nash equilibria.[34] Since the estimates are, however, in any case still close in magnitude to the relative frequency predicted by the Nash equilibria, it has been investigated whether they deviate from the Nash equilibria in a predictable manner.

Costa-Gomes and Weizsäcker (2008) and Fudenberg and Liang (2017) investigate first-period play, finding that players believe others to act like level-1 players in the initial round of a game. Costa-Gomes and Weizsäcker, however, find that there is a lot of stochasticity in behavior. The term "level 1" can be traced back to the literature on finite depth of reasoning, starting with Stahl (1993) and Nagel (1995),[35] and means that players give a (myopic) best response to random play. For example, in a 2x2 game level 1 thinking means believing that the opponent plays each action with a probability of 50 % and choosing a best-response according to that. A level-2 player, anticipating that agents may think that way, best responds to level-1 play etc.

I formulate a theory of noisy level-1 play. Consider for example the row player that needs to form beliefs about the initial probability of the column player playing left as opposed to right: $p_0^{COL}(L)$. Denote the Nash equilibrium $p^{COL,*}(L)$ and remember that in the Nash equilibrium, players mix their play so that every player (correctly) believes that she cannot predict her opponent. If players suspect to more likely observe a level-1 response from their opponents, this is no longer true. Without loss of generality, suppose further that the level-1 response of column is L. If the row player believes that her opponent most likely plays a level-1 response, she would expect that the column player "overplays" L with respect to equilibrium, so that $\hat{p}_0^{COL}(L) > p^{COL,*}(L)$. This prediction can be tested using the estimates of $\hat{p}_0^{COL}(L)$. For every estimate of $\hat{p}_0^{COL}(L)$, there are two possible outcomes: either the sign of the bias $\delta = \hat{p}_0^{COL}(L) - p^{COL,*}(L)$ is in the direction (> 0 or < 0) that level 1 would predict or it is not. If beliefs about initial play were unrelated to level 1, then about 50 % of the biases would be in the direction predicted by level 1 and 50 % would not be. This hypothesis can be tested using a one-sided binomial test. Using the estimates

[34]For Thompson Sampling, the p-value for the row player is < 0.05 except for games 9 (p-value: 0.2305) and 10 (p-value: 0.1406). For the column player, the p-value is < 0.01 in all games except for 3 (p-value: 1.0000), 6 (p-value: 0.2500), 7 (p-value: 0.3008) and 10 (p-value: 0.2500). For the Bayesian logit, the p-value for the row player is < 0.05 except in games 6 (p-value: 0.7383), 8 (p-value: 0.1289) and 9 (p-value: 0.3125). The p-value for the column player is < 0.01 in all games except for 7 (p-value: 0.0898) and 9 (p-value: 0.0547).

[35]See Nagel et al. (2016) for a historical survey of level k.

of all 90 subject pairs to provide sufficient power to the test, the null hypothesis can be rejected in favor of beliefs about the opponent being biased towards level-1 play for Thompson Sampling for both the row (p-value: 0.0000) and the column player (p-value: 0.0042). Similar results hold for the Bayesian logit (row: p-value: 0.0000; column: p-value: 0.0005). A purely Bayesian approach without any stochasticity would deterministically predict the row player in the above example to best-respond to level-1 play, being known as "level 2". However, a Thompson Sampler draws her estimate $\hat{p}_0^{COL}(L)$ from a beta distribution, which introduces randomness in her play. This is consistent with Costa-Gomes and Weizsäcker observing stochasticity in their experimental data.

Based on the above results, the following formal theory is proposed for the initial priors for the row player (analogously for the column player):

$$\alpha_0^{ROW} = \begin{cases} (p^{COL,*}(L) + \delta) \cdot N & \text{if level 1 predicts } a_0^{COL} = L \\ (p^{COL,*}(L) - \delta) \cdot N & \text{if level 1 predicts } a_0^{COL} = R \end{cases} \tag{40}$$

$$\beta_0^{ROW} = \begin{cases} (1 - (p^{COL,*}(L) + \delta)) \cdot N & \text{if level 1 predicts } a_0^{COL} = L \\ (1 - (p^{COL,*}(L) - \delta)) \cdot N & \text{if level 1 predicts } a_0^{COL} = R \end{cases} \tag{41}$$

$p^{COL,*}(L)$ is the probability of the column player playing L predicted by the Nash equilibrium. N the size of the "hypothetical" sample that players have in mind before playing. If N is high, the weight to accumulated experience during the game is low and players are very confident about their priors.

$\delta \geq 0$ is the bias from Nash equilibrium, incorporating the fact that players believe that their opponents might be level-1 thinkers. Suppose, without loss of generality, that the column player's level-1 response is L. If the row player believes that the column player most likely engages in level-1 thinking, it is reasonable for a player to believe that the opponent "overplays" L with respect to equilibrium, so that $(p^{COL,*}(L) + \delta)$.

TS Thompson Sampling has three free parameters corresponding to N, ξ, δ. N is the size of a "hypothetical" sample players have in mind before starting to play the game, ξ captures the weight given to a new signal. δ, being defined as $\hat{p}^{COL}(L) - \hat{p}^{COL,*}(L)$ for the row player (analogously for the column player $\hat{p}^{ROW}(T) - \hat{p}^{ROW,*}(T)$), is the amount by which the perceived initial probability of the opponent playing L (T) is biased towards level 1.

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 four free parameters: N, ξ, δ as in Thompson Sampling as well as the λ-parameter.[36]

[36] An unrestricted estimation chooses negative parameter values for δ in four out of the 10 estimations. Restricting the parameter values to be positive or $\delta = 0$ does not substantially change the out-of-sample predictions and only worsens the in-sample fit. Thus, the unrestricted estimation results are reported.