# Essays on Learning and Econometrics

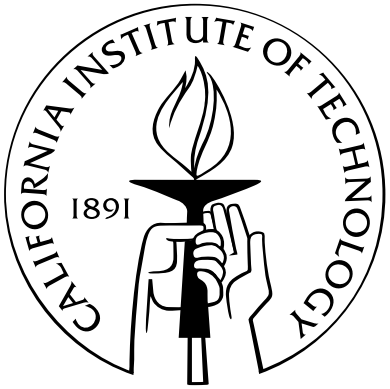
Thesis by

Yutaka Kayaba

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy




California Institute of Technology

Pasadena, California

2013

(Defended September 13, 2012)

To my family.

# Acknowledgements

I would like to express my deepest gratitude to my advisor, Peter Bossaerts, who allowed me to conduct an exciting, thoughtful project. His commitment, patience and devotion to my thesis are unparalleled, and without his contagious enthusiasm, this thesis would not be possible. I owe an equal amount of gratitude to the two other advisors, Matthew Shum and Colin Camerer. As a coauthor of the Chapter 2, Matt was always available to me and gave up countless hours to teach and guide me. Thanks to his effort, I successfully developed my essential skills as a scientist in the collaborative work with him. Not only providing me invaluable insights on behavioral economics, Colin also has taken care of me extensively, especially during and after the earthquake and tsunami in 2011. I sincerely appreciate his messages of heartfelt concerns during my mental turmoil right after the disaster, encouraging me to come back to my thesis work. I am also grateful to Yingyao Hu, Ben Gillen, Cary Frydman, Keise Izuma, Mathieu d'Acremont, Ian Krajbich, Mitsuru Igami and Taisuke Imai for their helpful contributions. Also I owe a great deal of thanks to Laurel Auchampaugh and Barbara Estrada for making my research smooth.

Last, but not the least, I am deeply indebted to my friends in my hometown, Sendai, who strived to surmount the unprecedented disaster in the earthquake and tsunami in 2011. I vividly remember that they all unite their efforts to save those who narrowly survived. Without their considerable and sustained efforts and patience, I might abandon my thesis to go back there to join.

This research has been in part supported by the Nakajima Foundation. I greatly acknowl-

edge their generous funding during the long academic journey.

# Abstract

This dissertation consists of two essays that focus on learning under state uncertainty and econometric applications for it, in which agents learn hidden state variables from their noisy measures sequentially.

In Chapter 2, "Nonparametric Learning Rules from Bandit Experiments: The Eyes Have It!", which is coauthored with Yingyao Hu and Matthew Shum, we assess, in a model-free manner, subjects' belief dynamics in a two-armed bandit learning experiment. A novel feature of our approach is to supplement the choice and reward data with subjects' *eye movements* during the experiment to pin down estimates of subjects' beliefs. Estimates show that subjects are more reluctant to "update down" following unsuccessful choices, than "update up" following successful choices. The profits from following the estimated learning and decision rules are smaller (by about 25% of typical experimental earnings) than what would be obtained from an fully rational Bayesian learning model, but comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning and a simple "win-stay" choice heuristic.

In Chapter 3, I examine the optimal learning models for predicting price dynamics under outlier risk. Two kinds of outlier risk in price processes are considered here; A price process in which outliers occur as its fundamental value has changed, and that with little fundamental change. In the latter process outliers occur as observation error, which are often referred as price anomalies in behavioral finance. The two optimal learning models are characterized with non-Gaussian Kalman filter as Bayesian reinforcement learning and

are solved numerically using sequential Monte Carlo sampling. Several key features are summarized in their learning rate and prediction error; The learning rate with outlier risk in fundamental value is a monotonically increasing function of absolute value of prediction error, while the learning rates with outlier risk in observation noise is a monotonically decreasing function. Interestingly, the uncertainly of the learning is seemingly identical among the two models, having a hump-shaped function of absolute value of prediction error.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

How optimal are humans' learning rules? The two essays in this thesis address the question on optimality of learning rules under state uncertainty, in which agents sequentially estimates hidden state variables from observable noisy signals of them. We present learning rules that explains the observed behavior in our laboratory experiments. The Chapter 2 presents a nonparametric estimates of subjects' learning rules in a dynamic two-armed bandit (probabilistic reversal-learning) task. Auxiliary measures of subjects' eye movements as they make their choices are employed to "pin down" subjects' beliefs in each round of the learning experiment. To our knowledge, the nonparametric estimation of learning rules is a new endeavor in both the behavioral learning literature, as well as the empirical literature in economics and marketing in which dynamic learning models are estimated structurally. While Chapter 2 is concerned with a nonparametric estimation of learning rule to assess subjects' optimality, in Chapter 3, I characterize an optimal learning under outlier risk, which has attracted considerable attention in finance recently (Taleb (2008)). In order to characterize the optimal learning rules under the outlier risk, I employ an unified approach of Bayesian learning and reinforcement learning, namely Bayesian reinforcement learning, in which the model components of reinforcement learning play crucial roles to describe the model characteristics.

A sizeable literature has been developed around structural estimation of empirical learning

models. Some representative papers include, Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005) and Chan and Hamilton (2006). This literature typically assumes that agents process information according to a forward-looking Bayesian learning model. This restrictive assumption is driven in part by data considerations: oftentimes, all that is observed are the sequences of agents' choices, so that a lot of (parametric) structure must be placed on the learning model for identification.

In controlled experimental settings, richer data are observed: not only subjects' choices, but also the outcomes (rewards) from their choices. In addition, there is also the opportunity to observe "auxiliary" measures of subjects' beliefs (or valuations), such as brain activity (cf. Boorman, Behrens, Woolrich, and Rushworth (2009), Hsu, Bhatt, Adolphs, Tranel, and Camerer (2005), Preuschoff, Bossaerts, and Quartz (2006) in the recent fMRI neuroscience literature) or attension measures, or "lookups", from mouse tracking (Camerer, Johnson, Rymon, and Sen (1993), Johnson, Camerer, Sen, and Rymon (2002), Costa-Gomes, Crawford, and Broseta (2001), Costa-Gomes and Crawford (2006), Gabaix, Laibson, Moloche, and Weinberg (2006), Brocas, Carrillo, Wang, and Camerer (2009)), or from eye tracking (Wang, Spezio, and Camerer (2010), Knoepfle, Wang, and Camerer (2009)).

Because of this additional data richness, researchers in the behavioral/experimental literature have been able to consider more flexible learning rules, and to test the fully rational Bayesian learning benchmark versus boundedly rational, non-Bayesian "reinforcement learning" rules (cf. Sutton and Barto (1998)). An incomplete list of papers which consider these questions includes Grether (1992), El-Gamal and Grether (1995), Nyarko and Schotter (2002), Charness and Levin (2005), Kuhnen and Knutson (2008), Camerer and Ho (1999), and Payzan-LeNestour and Bossaerts (2011). Particularly, reinforcement learning has attracted considerable attention in the recent neuroeconomics and decision neuroscience literature (cf. Glimcher, Camerer, Poldrack, and Fehr (2008), Rushworth and Behrens (2008)), ever since studies showing that the "prediction errors" of these models are apparently encoded in certain regions of the brain (cf. Schultz, Dayan, and Montague (1997)) for evidence from primates). Recently, reinforcement learning models have also been used to

explain some observed anomalies in savings and investment behavior (e.g., Choi, Laibson, Madrian, and Metrick (2009), Strahilevitz, Odean, and Barber (2011)).

In Chapter 2, we take a new approach to assessing learning in experimental settings. Taking advantage of recent developments in the econometrics of estimating dynamic models with serially correlated unobservables, we use the observed experimental and auxiliary data, namely eye-tracking data, to estimate, nonparametrically, subjects' choice probabilities and learning rules, without imposing *a priori* functional forms on these functions. Thus, our learning rules can be reasonably interpreted as "what the subjects actually think", as reflected in their observed choices. Subsequently, we compare our estimated learning rules to specific parameterized learning rules which have been considered in the previous literature, including the Bayesian and reinforcement-learning models.

While Chapter 2 is concerned with nonparametric estimation of learning rule compared to Bayesian learning and reinforcement learning, in Chapter 3, I characterize optimal learning rules in a unified approach of Bayesian learning and reinforcement learning, especially focusing on financial applications to learning hidden states under outlier risk (Taleb (2008)). Here agents learn dynamics of hidden state variables from observable signals in financial markets with outlier risk, where the former is unobservable fundamental values and the latter observable prices. Two kinds of outlier risk in the price processes are considered here; A price process in which outliers occur as its fundamental value has changed, and that with little fundamental change. In the latter process outliers occur as observation error, which are often referred as price anomalies in behavioral finance. The two optimal learning models are characterized with non-Gaussian Kalman filter as Bayesian reinforcement learning and are solved numerically using sequential Monte Carlo sampling. To characterize the optimal, Bayesian learning under the outlier risk, two model components of reinforcement learning —prediction error and learning rate —play crucial roles.

Chapter 3 is a part of broader, interdisciplinary research agenda among finance, machine learning and decision neuroscience, in which we explore human nature of predicting price

dynamics under outlier risk in neurofinance. Recent studies in decision neuroscience (Yu and Dayan (2005), Aston-Jones and Cohen (2005)) and psychology (Kruschke and Johansen (1999)) revealed that salient information often alter mental models in the human mind, suggesting that humans are more likely to see fundamental changes behind salient, large price changes. However as behavioral finance argues, in the modern financial market, price dynamics often shows salient outliers that are seemingly unrelated to fundamental changes. We speculate that the counterintuitive contradiction between humans' impulsive thinking behind outliers and the price anomalies in financial market could be a source of humans' suboptimal behavior in price predictions in the financial markets. Currently less is known about the neurobiological foundations of information processing of outliers in human minds, however, the comprehensive studies of the issue would help us uncover the nature of human suboptimal behavior in the financial markets.

# Chapter 2

# Nonparametric Learning Rules from Bandit Experiments: The Eyes Have It!

## 2.1 Introduction

How do individuals learn from past experience in dynamic choice environments? A growing literature has documented, using both experimental and field data, that the benchmark fully rational Bayesian learning model appears deficient at characterizing actual decision-making in real-world settings.[1] Other papers have demonstrated that observed choices in strategic settings with asymmetric information are typically not consistent with subjects' having Bayesian (equilibrium) beliefs regarding the private information of their rivals.[2] Recently, non-Bayesian *reinforcement learning* (Sutton and Barto (1998)) models have also been used to explain some observed anomalies in savings and investment behavior (e.g., Choi, Laibson, Madrian, and Metrick (2009), Strahilevitz, Odean, and Barber (2011)).

[1] An incomplete list of papers which consider these questions includes Grether (1992), El-Gamal and Grether (1995), Nyarko and Schotter (2002), Charness and Levin (2005), Kuhnen and Knutson (2008), and Payzan-LeNestour and Bossaerts (2011).

[2] For instance, Bajari and Hortaçsu (2005), Goldfarb and Xiao (2011), Ho and Su (2010), Aguirregabiria and Magesan (2011), Crawford and Iriberri (2007), Gillen (2011), Brown, Camerer, and Lovallo (2010), Brocas, Carrillo, Wang, and Camerer (2009) .

Given the lack of consensus in the literature about what the actual learning rules used by agents in real-world decision environments are, there is a need for these rules to be estimated in a manner flexible enough to accommodate alternative models of learning. In this paper, we propose a new approach for assessing agents' belief dynamics. In an experimental setting, we utilize data on subjects' *eye movements* during the experiment to aid our inference regarding the learning (or belief-updating) rules used by subjects in their decision-making process. Previous studies (e.g., Shimojo, Simion, Shimojo, and Scheier (2003)) have established a connection between subjects' eye movements and gaze durations and their valuations in choice experiments. We exploit this connection and use gaze durations to *pin down* subjects' evolving beliefs in a dynamic choice setting.

Eye-movement data are an example of novel non-choice variables, the measurement and analysis of which has constituted an important strand in experimental economics. Caplin and Dean (2008) broadly call these auxiliary measures "neuroeconomic" data, in the sense of data other than the usual choice and rewards data gathered from typical experiments. Besides eye tracking (Krajbich, Armel, and Rangel (2010), Wang, Spezio, and Camerer (2010), Knoepfle, Wang, and Camerer (2009)), other examples of such data include measurements of brain activity (Boorman, Behrens, Woolrich, and Rushworth (2009), Hsu, Bhatt, Adolphs, Tranel, and Camerer (2005), Preuschoff, Bossaerts, and Quartz (2006)), pupil dilation (Preuschoff, Marius Hart, and Einhauser (2011)), skin conductance response (Sokol-Hessner, Hsu, Curley, Delgado, Camerer, and Phelps (2009)) and mouse tracking (Camerer, Johnson, Rymon, and Sen (1993), Johnson, Camerer, Sen, and Rymon (2002), Costa-Gomes, Crawford, and Broseta (2001), Costa-Gomes and Crawford (2006), Brocas, Carrillo, Wang, and Camerer (2009), Gabaix, Laibson, Moloche, and Weinberg (2006)).

To our knowledge, we are the first to use such "neuroeconomic" data in estimating behavioral decision-making models.[3] Taking advantage of recent developments in the economet-

[3]The existing neuroeconomic literature has used neuroeconomic data to either test or select among behavioral models (cf. Caplin, Dean, Glimcher, and Rutledge (2010); Symmonds, Bossaerts, and Dolan (2010); Wunderlich, Beierholm, Bossaerts, and O'Doherty (2011); Hsu, Bhatt, Adolphs, Tranel, and Camerer (2005); Hampton, Bossaerts, and O'Doherty (2008); Hsu, Krajbich, Zhao, and Camerer (2009).) but not,

rics of dynamic measurement error models, we use the observed choice and eye- tracking data to estimate subjects' decision rules and learning rules, without imposing a priori functional forms on these functions. Estimating the learning rules in such a model-free manner allows us to assess the optimality of subjects' choices in learning experiments in a manner quite distinct from that taken in the existing literature.

Our main results are as follows. First, our estimated learning rules do not correspond to any one of the existing learning models. Rather, we find that beliefs are reward-asymmetric, in that subjects are more reluctant to "update down" following unsuccessful (low-reward) choices, than "update up" following successful (high-reward) choices. Such asymmetries are novel relative to existing learning models (such as reinforcement or Bayesian learning); moreover, from a payoff perspective, they are suboptimal relative to the fully rational Bayesian benchmark.

Correspondingly, we find that, using the estimated learning rules, subjects' payoffs are, at the median, \$4 (or about two cents per choice) lower than under the Bayesian benchmark; this difference represents about 25% of typical experimental earnings (not including the fixed show-up fee). However, subjects' payoffs under the estimated choice and learning rules are comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning.

In the next section, we describe the dynamic two-armed bandit learning (probabilistic reversal-learning) experiment, and the eye-movement data gathered by the eye-tracker machine. In Section 3, we present an econometric model of subjects' choices in the bandit model, and discuss nonparametric identification and estimation. In Section 4, we describe the experimental data, and present our nonparametric estimates of subjects decision rules and learning rules. Section 5 contains a comparison of our estimated learning rules to "standard" learning rules, including those from the Bayesian and non-Bayesian reinforcement-learning models. Section 6 concludes.

---

as far as we are aware, for estimating models.

## 2.2 Two-armed bandit "reversal-learning" experiment

Our experiments are adapted from the "reversal-learning" experiment used in Hampton, Bossaerts, and O'Doherty (2006). In the experiments, subjects make repeated choices between two actions (which we call interchangeably "arms" or "slot machines" in what follows): in trial $t$, the subject chooses $Y_t \in \{1(=\text{"green"}), 2(=\text{"blue"})\}$. The rewards generated by these two arms are changing across trials, as described by the state variable $S_t \in \{1, 2\}$, which is never observed by subjects. When $S_t = 1$, then green (blue) is the "good" ("bad") state, whereas if $S_t = 2$, then blue (green) is the "good" ("bad") state.

The rewards $R_t$ that the subject receives in trial $t$ depends on the action taken, as well as (stochastically) on the current state: the reward process is

$$R_t = \begin{cases} \pm\$0.50 \text{ with prob. } 50\% \pm 20\% & \text{if good arm chosen} \\ \pm\$0.50 \text{ with prob. } 50\% \mp 10\% & \text{if bad arm chosen.} \end{cases} \tag{2.1}$$

For convenience, we use the notation $R_t = 1$ to denote the negative reward (-\$0.50), and $R_t = 2$ to denote the positive reward (\$0.50).

The state evolves according to an exogenous binary Markov process. At the beginning of each block, the initial state $S_1 \in \{1, 2\}$ is chosen with probability 0.5, randomly across all subjects and all blocks. Subsequently, the state evolves with transition probabilities[4]

| $P(S_{t+1}\|S_t)$ | $S_t = 1$ | $S_t = 2$ |
|---|---|---|
| $S_{t+1} = 1$ | 0.85 | 0.15 |
| $S_{t+1} = 2$ | 0.15 | 0.85 |

. (2.2)

[4]This aspect of our model differs from Hampton, Bossaerts, and O'Doherty (2006), who make the non-Markovian assumption that the state $S_t$ changes with probability 25% after a subject has chosen the good arm four successive times. Estimating such non-Markovian models would, typically, require including another state variable, which describe how uncertain the subject is at any point in time about the underlying state (such as the variance of rewards). In principle, our estimation method can be extended to allow for these additional state variables, but since we take a nonparametric approach, a much larger sample size (far beyond typical samples sizes in experimental work) would be required to obtain reasonable estimates.

Because $S_t$ is not observed by subjects, and is serially correlated over time, subject have an opportunity to learn and update their beliefs about the current state on the basis of past rewards. Moreover, because $S_t$ changes randomly over time, so that the identity of the good arm varies across trials, this is called a "probabilistic reversal-learning" experiment.

### 2.2.1 Optimal decision-making in reversal-learning model

Before proceeding to describe the experimental data, we consider how subjects should optimally make decisions in the dynamic reversal-learning model used in our experiments. The qualitative features of the optimal decision and belief-updating rules presented here will motivate the assumptions which underlie the empirical learning model which we estimate in this paper. As in the experiments, we consider a finite (25 period) dynamic optimization problem, in which each subject aims to choose a sequence of actions to maximize expected rewards $\mathbb{E}\left[\sum_{t=1}^{25} R_t\right]$. (The details of this model are given in Appendix A.1.)

Let $B_t^*$ denote the probability (given by Bayes' Rule) denote the probability that a subject places on "green" being the good arm in period $t$, conditional on the whole experimental history up to then. We evaluate the optimal decision rules—the mapping from period $t$ beliefs $B_t^*$ to a period $t$ choice—in this dynamic Bayesian learning model by computer simulation. Importantly, we accommodate nonstationarity in the problem, in that our simulations allow the decision rules to differ arbitrarily across periods. This permits the relationship between subjects' choices and their beliefs $B_t^*$ to vary across periods, depending perhaps on the periods remaining in the experiment, or to allow for history dependence in either choices or the belief-updating rule. An important maintained assumption in this paper is that subjects' decision rules are solely a function of the current state probabilities $B_t^*$, so that by allowing the decision rules to vary across periods in these simulations, we can assess the restrictiveness of such an assumption.

The important qualitative features of optimal decision-making are summarized in the optimal decision rules, which we plot in Figure 2.2 for four periods $t = 1,\ 10,\ 20,\ 25$. Two

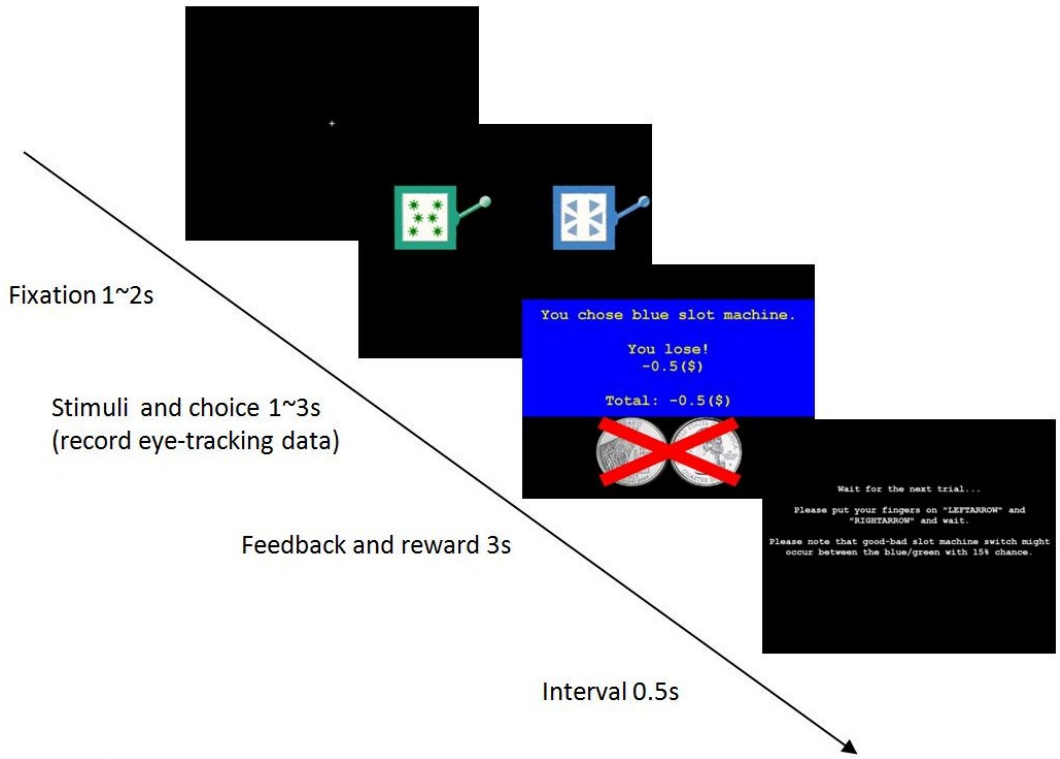
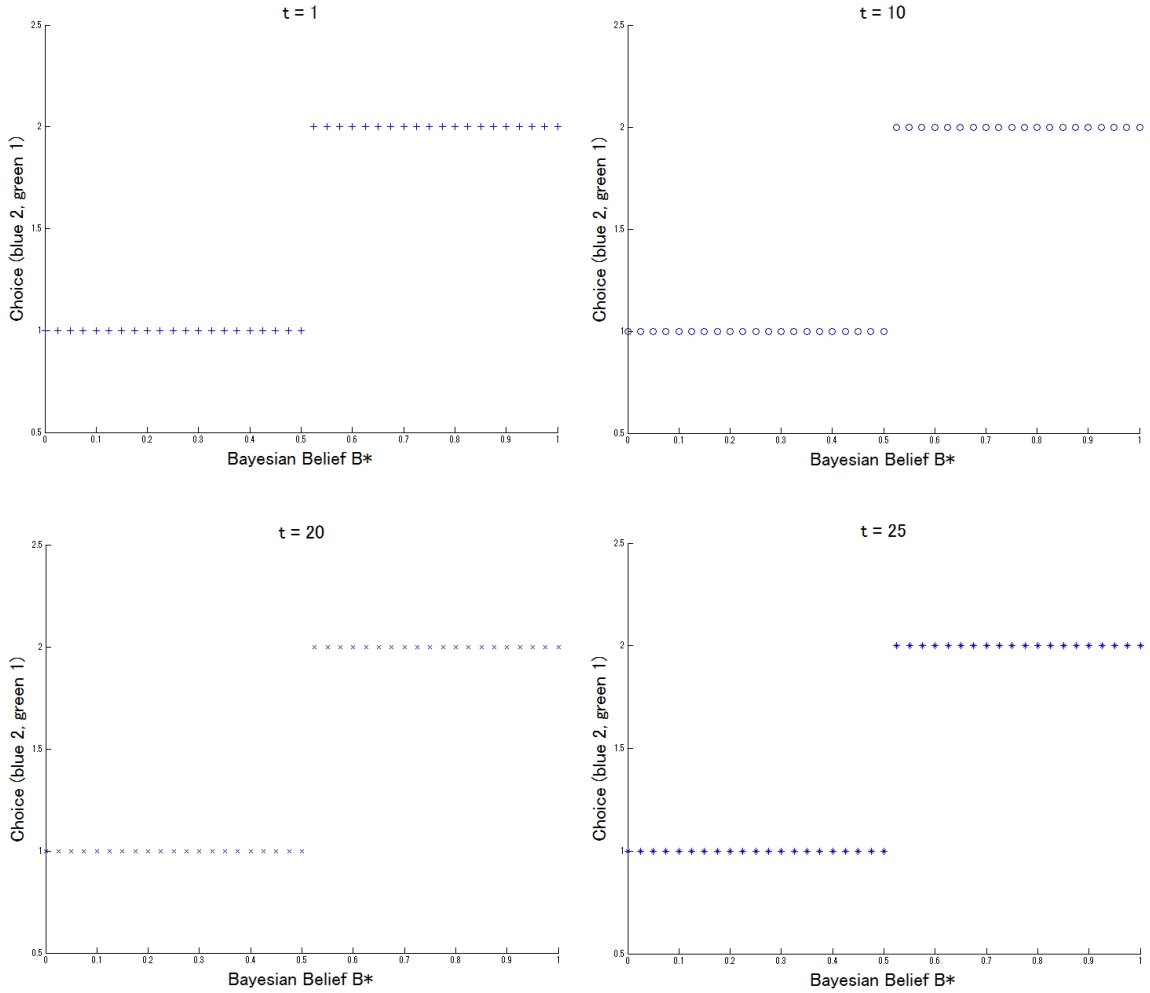



Figure 2.1: Timeline of a trial

After a fixation on the cross (top screen), two slot machines are presented (the left-right position is randomized; second screen). Subjects' eye movements are recorded by the eye-tracking machine here. Subjects choose by pressing the left (right) arrow key to indicate a choice of the left (right) slot machine. After choosing (third screen), a positive reward (depicted by two quarters) or negative reward (two quarters covered by a red X) is delivered, along with feedback about the subject's choice highlighted against a background color corresponding to the choice. In the bottom screen, a subject is transitioned to the next trial, and reminded that the a slot machine may switch from "good" to "bad" (and vice versa) with probability 15%.

Figure 2.2: Optimal decision rules in reversal-learning model

Plotted for periods $t = 1, 10, 20, 25$.




x-axis: Bayesian belief $B_t^*$; y-axis: action choice

features are apparent. First, we see that the decision rules are identical across all the periods, indicating that they are *stationary*. Second, the optimal decision rule takes a simple form: in each period, the subject chooses simply blue once the current belief that the blue arm is "good" exceeds 50%. This is a *myopic* decision rule.

Both of these features—stationarity and myopia of decision rules—are specific to the reversal-learning setup considered here, and differs in important ways from optimal decision-making in the standard multi-armed bandit (MAB) problem (cf. Jones and Gittins (1972), Banks and Sundarum (1992)), in which the states of the bandits are fixed over all periods and the bandits are "independent" in that a reward from one bandit is uninformative about the state of another bandit. The optimal Bayesian decision rule in the standard MAB model features exploration (or "experimentation"), which recommends sacrificing current rewards to achieve longer-term payoffs; this makes simple myopic decision-making (choosing the bandit which currently has the higher expected reward) suboptimal. In the reversal-learning setting, however, the states of the bandits are negatively related, so that positive information about one slot machine implies negative information about the other. Apparently, as shown by these optimal decision rules, this eliminates most of the incentives for subjects to experiment.

Moreover, in a finite-horizon decision environment, such as the experiments considered here, the value of information decreases exogenously as the final period approaches, resulting in reduced incentives for experimentation; this implies a nonstationary decision rule. Under reversal learning, however, the lack of experimentation leads to stationary decision rules, even in a finite-horizon problem.

### 2.2.2 Experimental data: preliminary analysis

The experiments were run over several weeks in November-December 2009. We used 21 subjects, recruited from the Caltech Social Science Experimental Laboratory (SSEL) subject pool consisting of undergraduate/graduate students, post-doctoral students, and community

members,[5] each playing for 200 rounds (broken up into 8 blocks of 25 trials). Most of the subjects completed the experiment within 40 minutes, including instruction and practice sessions. Subjects were paid a fixed show-up fee ($20), in addition to the amount won during the experiment, which was $14.20 on average.[6]

Subjects were informed of the reward structure for good and bad slot machines, and the Markov transition probabilities for state transitions (reversals), but were not informed which state was occurring in each trial. In Figure 2.1, we present the time line and some screenshots from the experiment. In addition, while performing the experiment, the subjects were attached to an eye-tracker machine, which recorded their eye movements. From this, we constructed the auxiliary variable $\tilde{Z}_t$, which measures the fraction of the reaction time (the time between the onset of a new round after fixation, and the subject's choice in that round) spent gazing at the picture of the "blue" slot machine on the computer screen.[7]

For each subject, and each round $t$, we observe the data $(Y_t, S_t, R_t, Z_t)$. Table 2.1 presents some summary statistics of the data. The top panel shows that, across all subjects and all trials, "green" (2108 choices) and "blue" (2092 choices) are chosen in almost-equal proportions. Moreover, from the second panel, we see that subjects obtain the high reward with frequency of roughly 57% ($\approx 2398/(2398+1802)$). This is slightly higher than, but significantly different from, 55%, which is the frequency which would obtain if the subjects were choosing completely randomly.[8] Hence, subjects appear to be "trying", which motivates our analysis of their learning rules. On the other hand, simulation of the optimal Bayesian decision rules (discussed above) show that the success rate from using the optimal decision

[5]Community members consisted of spouses of students at either Caltech or Pasadena City College (a two-year junior college). While the results reported below were obtained by pooling the data across all subjects, we also estimated the model separately for the subsamples of Caltech students, vs. community members. There were few noticeable differences in the results across these classes of subjects.

[6]For comparison, purely random choices would have earned $10 on average.

[7]Across trials, the location of the "blue" and "green" slot machines were randomized, so that the same color is not always located on the same side of the computer screen. This controls for any "right side bias" which may be present (see discussion further below).

[8]This is the marginal probability of a good reward, which equals $0.5(0.7+0.4)$ from Eq. (2.1). The t-statistic for the null that subjects are choosing randomly equals 169.67, so that hypothesis is strongly rejected.

Table 2.1: Summary statistics for experimental data

| | 1(green) | 2(blue) |
|---|---|---|
| $Y$: subjects' choices | 2108 | 2092 |

| | 1 ($0.50) | 2 (-$0.50) |
|---|---|---|
| $R$: rewards | 2398 | 1802 |

| | mean | median | upper 5% | lower 5% |
|---|---|---|---|---|
| $\widetilde{Z}$: eye movement measure[a] | -0.0309 | 0 | 1.3987 | -1.4091 |
| $RT$: reaction time ($10^{-2}$ secs) | 88.22 | 59.3 | 212.2 | 36.8 |

[a]Defined in Eq. (2.3)

rule is only 58.4%, which is just slightly higher than the in-sample success rate found in the experiments. It appears, then, that in the reversal-learning setting, the success rate intrinsically varies quite narrowly between 55% and 58.4%.

In Table 2.2, we present the conditional probabilities of choices in period $t$, conditional on choices and rewards from the previous period $(Y_t|Y_{t-1}, R_{t-1})$. This can be interpreted as a "reduced-form" decision rule for the subjects. The top row in that table contains the reduced-form probabilities of choosing the green arm. Looking at the second (fourth) entry in this row, we see that after a successful choice of green (blue), a subject replays this strategy with probability 0.86 (0.88=1–0.12). Thus subjects appear to replay successful strategies, corresponding to a "win-stay" rule-of-thumb.

However subjects appear reluctant to give up *unsuccessful* strategies. The probability of replaying a strategy after an unsuccessful choice of the same strategy is around 50% for both the blue and green choices (ie. the first and third entries in this row). Thus, subjects tend to randomize after unsuccessful strategies. As far as we are aware, such an "asymmetric" choice rule is new in the literature; moreover, as we will see below, this is echoed in the

"asymmetric" belief-updating rule which we estimate.

In the remainder of Table 2.2, we also present the same choice probabilities, calculated for each subject individually. There is some degree of heterogeneity in subjects' strategies. Looking at columns 2 and 4 or the table, we see that, for the most part, subjects pursue a "win-stay" strategy: the probabilities in the second column are mainly $>> 50\%$, and those in the fourth column are most $<< 50\%$. However, looking at columns 1 and 3, we see that there is significant heterogeneity in subjects' choices following a low reward. In these cases, randomization (which we classify as a choice probability between 40–60%) is the modal strategy among subjects; strikingly, however, a number of subjects continue replaying an unsuccessful strategy: for example, subjects 3, 8, and 11 continue to choose "green" with probabilities of 79%, 89%, and 79% even after a previous choice of green yielded a negative reward.[9]

One common feature of the choice strategies across all subjects is that choices are serially correlated across periods, conditional on rewards. This serial correlation is very informative for identifying the beliefs $X_t^*$. Essentially, we present a model below in which serial correlation in choices across periods arises due to beliefs—thus, beliefs (which are unobserved to the researcher) are the reason for serial correlation of choices. We will discuss this in more detail in the next section, in which the empirical model is presented formally.

### 2.2.3 Remarks on eye-tracking measure

Although "lookups" data from mouse tracking has been employed in several studies to understand cognitive processes in economic decisions[10], recently, eye tracking has been

[9] In the reversal-learning model, however, such a strategy is not obviously irrational; because the identity of the good arm changes exogenously across periods, an arm that was bad last period (i.e., yielding a low reward) may indeed be good in the next period.

[10] Studies by Camerer, Johnson, Rymon, and Sen (1993) and Johnson, Camerer, Sen, and Rymon (2002) (as well as the development of "Mouselab" system and its application to choice tasks in Payne, Bettman, and Johnson (1993)) were the pioneering work in economics in which they applied mouse tracking in alternating-offer bargaining games. Then, the attension measures were employed to study forward induction (Camerer and Johnson (2004)), level-k models in matrix-games (Costa-Gomes, Crawford, and Broseta (2001), two-

Table 2.2: "Reduced-form" decision rule: $P(Y_t = 1(green)|Y_{t-1}, R_{t-1})$

| $(Y_{t-1}, R_{t-1})$: | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| All Subjects: | 0.5075 | 0.8652 | 0.5089 | 0.1189 |
| | (0.0169) | (0.0094) | (0.1169) | (0.0090) |
| $Subject1$: | 0.1799 | 0.5192 | 0.8128 | 0.364 |
| | (0.0655) | (0.0684) | (0.0595) | (0.0603) |
| $Subject2$: | 0.1051 | 0.9820 | 0.9449 | 0 |
| | (0.0498) | (0.0171) | (0.0381) | (0) |
| $Subject3$: | 0.7938 | 0.9859 | 0.3340 | 0 |
| | (0.0591) | (0.0136) | (0.0871) | (0) |
| $Subject4$: | 0.3244 | 0.8796 | 0.6492 | 0.0610 |
| | (0.0704) | (0.0514) | (0.0726) | (0.0283) |
| $Subject5$: | 0.0419 | 0.8796 | 0.6492 | 0.0610 |
| | (0.0292) | (0.0236) | (0.0325) | (0.0461) |
| $Subject6$: | 0.2570 | 0.7498 | 0.8159 | 0.2021 |
| | (0.0652) | (0.0592) | (0.0602) | (0.0532) |
| $Subject7$: | 0.5792 | 0.9242 | 0.4647 | 0.0796 |
| | (0.0751) | (0.0371) | (0.0731) | (0.0379) |
| $Subject8$: | 0.8931 | 0.9803 | 0.1013 | 0.0165 |
| | (0.0496) | (0.0186) | (0.0482) | (0.0163) |
| $Subject9$: | 0.6377 | 1.0000 | 0.2741 | 0 |
| | (0.0831) | (0) | (0.0655) | (0) |
| $Subject10$: | 0.1986 | 0.9344 | 0.8037 | 0 |
| | (0.0622) | (0.0352) | (0.0587) | (0) |
| $Subject11$: | 0.7859 | 1.0000 | 0.4306 | 0 |
| | (0.0575) | (0) | (0.0870) | (0) |
| $Subject12$: | 0.5883 | 0.9262 | 0.3741 | 0.0131 |
| | (0.0841) | (0.0406) | (0.0733) | (0.0129) |
| $Subject13$: | 0.6741 | 0.8907 | 0.1962 | 0.2085 |
| | (0.0705) | (0.0462) | (0.0581) | (0.0539) |
| $Subject14$: | 0.4730 | 0.6147 | 0.5363 | 0.3842 |
| | (0.0831) | (0.0653) | (0.0735) | (0.0664) |
| $Subject15$: | 0.6759 | 0.9789 | 0.3351 | 0 |
| | (0.0761) | (0.0206) | (0.0714) | (0) |
| $Subject16$: | 0.4595 | 0.9135 | 0.5443 | 0.1953 |
| | (0.0715) | (0.0316) | (0.0742) | (0.0666) |
| $Subject17$: | 0.6358 | 0.5202 | 0.5322 | 0.4644 |
| | (0.0660) | (0.0706) | (0.0780) | (0.0748) |
| $Subject18$: | 0.6333 | 1.0000 | 0.2901 | 0 |
| | (0.0834) | (0) | (0.0734) | (0) |
| $Subject19$: | 0.6144 | 0.8197 | 0.5808 | 0.2013 |
| | (0.0702) | (0.0444) | (0.0806) | (0.0625) |
| $Subject20$: | 0.3699 | 0.5741 | 0.3699 | 0.3554 |
| | (0.0858) | (0.0707) | (0.0665) | (0.0621) |
| $Subject21$: | 0.6990 | 0.9602 | 0.2934 | 0.0177 |
| | (0.0658) | (0.0274) | (0.0693) | (0.0171) |

Note: standard errors (in parentheses) computed using 1000 bootstrap resamples

employed to measure eye fixation more precisely[11] in various decision environments: to determine how subjects detect truth-telling or deception in sender-receiver games (Wang, Spezio, and Camerer (2010)); how consumers evaluate comparatively a large number of commodities, as in a supermarket setting (Reutskaja, Nagel, Camerer, and Rangel (2011)); and the relationship between visual attention (as measured by eye-fixations) and choices of commodities in choice tasks (cf. Krajbich, Armel, and Rangel (2010), Armel and Rangel (2008), Armel, Beaumel, and Rangel (2008), Rangel (2009)).[12]

Our use of eye movements in this paper is predicated on an assumption that gaze is related to beliefs of expected rewards. This is motivated by some recent results in behavioral neuroscience. Shimojo, Simion, Shimojo, and Scheier (2003) studied this in binary "face choice" tasks, in which subjects are asked to choose one of the two presented faces on the basis of various criteria. (Our two-armed bandit task is very similar in construction.) These authors find that, when subjects are asked to choose a face based on attractiveness, their eye movements tend to be directed to the preferred face as time goes, which they name as "gaze cascading effect". Interestingly, the relationship between gaze direction and the chosen face becomes significantly weaker when subjects are asked to choose a face based on shape and "unattractiveness". This strongly suggests that directed gaze duration reflects preferences, rather than choices.

This work echoes primate experiments reported in Lauwereyns, Watanabe, Coe, and Hikosaka (2002) and Kawagoe, Takikawa, and Hikosaka (1998) (see the survey in Hikosaka, Nakamura, and Nakahara (2006)), which shows that primates tend to direct their gaze at locations where rewards are available. They also establish a physiological basis for this relationship,

person guessing games (Costa-Gomes and Crawford (2006)), two-person games with private information about payoff-relevant states (Brocas, Carrillo, Wang, and Camerer (2009), a boundedly rational, directed cognition model in goods choice tasks (Gabaix, Laibson, Moloche, and Weinberg (2006)).

[11] Most of the mouse-tracking studies above employed "Mouselab" system in which subjects have to move a cursor into a box to open its contents. However, as Wang, Spezio, and Camerer (2010) argued, "experimenter cannot be certain the subject is actually looking at (and processing) the contents of the open box," which motivated them to use eye tracking in order to measure eye fixation more reliably.

[12] Eye tracking has also been used in marketing studies to evaluate the relationship between visual attention to advertisements (e.g.,Lohse (1997)) and subsequentsales of advertised items (e.g., Zhang, Wedel, and Pieters (2009)).

by showing a connection between eye movements and reward-sensitive neuronal activities in the basal ganglia part of the brain. These results, which link gaze direction and duration with expected rewards, provide some precedence and justification to our use of eye movements as "noisy measures" of beliefs, which likewise reflect perceptions of expected rewards from the slot machines.[13]

Based on the papers above, we define $\tilde{Z}_{it}$, our raw eye-movement measure, as the difference in the gaze duration directed at the blue and green slot machines, normalized by the total reaction time:

$$\tilde{Z}_t = (Z_{b,t} - Z_{g,t})/RT_t; \tag{2.3}$$

that is, for trial $t$, $Z_{b(g),t}$ is the fixation duration at the blue (green) slot machine, and $RT_t$ is the reaction time, i.e., the time between the onset of the trial after fixation, and the subject's choice.[14] Thus, $\tilde{Z}_t$ measures how much longer a subject looks at the blue slot machine than the green one during the $t$-th trial, with a larger (smaller) value of $\tilde{Z}_t$ implying longer fixation time at the blue (green) slot machine. Summary statistics on this measure are given in the bottom panel of Table 2.1. There, we see that the average reaction time is 0.88 seconds, and that the median value of $\tilde{Z}_t$ is zero, implying an equal amount of time directed to each of the two slot machine.[15]

[13] An alternative to using eye movements to proxy for beliefs would have been to elicit beliefs (as in Nyarko and Schotter (2002)). However, given the length of our experiments (8 trials of 25 periods each), and our need to have beliefs for each period, it seemed infeasible to elicit beliefs. Indeed, in our pilot experiments, we tried eliciting beliefs randomly after some periods, and found that this made the experiments unduly long.

[14] Furthermore, in order to control for subject-specific heterogeneity, we normalize $\tilde{Z}_t$ across subjects by dividing by the subject-specific standard deviation of $\tilde{Z}_t$, across all rounds for each subject.

[15] Following the suggestions of a referee, we also considered an alternative definition of the eye-movement measure $\tilde{Z}_t = (Z_{b,t} - Z_{g,t})/(Z_{b,t} + Z_{g,t})$, in which the time spent gazing at the middle of the screen (which is $RT_t - Z_{b,t} - Z_{g,t}$) is not included in the denominator. This allows for the possibility that the time spent gazing in the middle may be indicative of "contemplation", and may lead to stronger subsequent beliefs. We found that the estimation results for the choice probabilities and learning rules from this alternative specification (which are available from the authors upon request) are quite similar to the results from our standard specification, which are reported below. This suggests that the time spent gazing in the middle of the screen is not that informative about the evolution of subjects' beliefs. Similarly, another referee suggested that the absolute reaction time $RT_t$ itself could be included in the definition of eye movements. However, we found that the absolute value of $\tilde{Z}_t$ is inversely related to $RT_t$; this suggests that our measure $\tilde{Z}_t$ appears to capture or contain the information in reaction time.

Figure A.2 in the Appendix A.6 contains the scatter plot of $Z_{b,t}$ versus $Z_{g,t}$. In our empirical work, we will discretize the eye-movement measure $\tilde{Z}_t$; to avoid confusion, in the following we use $\tilde{Z}_t$ to denote the undiscretized eye-movement measure, and $Z_t$ the discretized measure, which we describe below.

## 2.3 Empirical model

In this section, we describe our econometric model of dynamic decision-making in the two-armed bandit experiment described above, and also discuss the identification and estimation of this model. Importantly, most of the crucial assumptions of the model are motivated by the structure of the optimal decision rules and learning (belief-updating) rules, as described in Section 2.2.1. That is, we do not consider the whole gamut of learning models here, but restrict attention to models which are "close" to optimal in that the structure of the learning and decision rules are the same as in the optimal model; however, the rules themselves are allowed to be different.

We introduce the variable $X_t^*$, which denotes the agent's round $t$ beliefs about the current state $S_t$; obviously, agents know their beliefs $X_t^*$, but these are unobserved by the researcher.[16] In what follows, we assume that both $X^*$ and $Z$ are discrete, and take support on $K$ distinct values which, without loss of generality, we denote $\{1, 2, \ldots, K\}$. We make the following assumptions regarding the subjects' learning and decision rules:

**Assumption 1** *Subjects' choice probabilities $P(Y_t|X_t^*)$ only depend on current beliefs. Moreover, the choice probabilities $P(Y_t = y|X_t^*)$ varies across different values of $X_t^*$ (i.e., beliefs affect actions).*

Because we interpret the unobserved variables $X_t^*$ here as a measurement of subjects' *current*

[16] $X_t^*$ corresponds to the prior beliefs $p_t$ from the previous section except that, further below, we will discretize $X_t^*$ and assume that it is integer-valued. Therefore, to prevent any confusion, we will use distinct notation $p_t$, $X_t^*$ to denote, respectively, the beliefs in the theoretical vs. the empirical model.

beliefs regarding which arm is currently the "good" one, the choice probability $P(Y_t|X_t^*)$ can be interpreted as that which arises from a "myopic" choice rule. As we remarked before, in Section 2.2.1, such an interpretation is justified by the simulation of the optimal choice rules under the reversal-learning setting, which showed that these rules are myopic and depend only on current beliefs.

Furthermore, Assumption 1 embodies an important exclusion restriction that, conditional on beliefs $X_t^*$, the observed action $Y_t$ is independent of the eye movement $Z_t$. As we will see below, this is a critical identification assumption which pins down the beliefs $X_t^*$ in the empirical model.

**Assumption 2** *The law of motion for $X_t^*$, which describes how subjects' beliefs change over time given the past actions and rewards, is called the* ***learning rule****. This is a controlled first-order Markov process, with transition probabilities $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$.*

This assumption is motivated by the structure of the optimal Bayesian belief-updating rule (cf. Eq. (A.1) in Appendix A.1), in which the period $t$ beliefs depend only on the past beliefs, actions, and rewards in period $t-1$. However, we allow the exact form of the learning rule to deviate from the exact Bayes formula.

**Assumption 3** *The eye-movement measure $Z_t$ is a noisy measure of beliefs $X_t^*$:*
*(i) Eye movements are serially uncorrelated conditional on beliefs: $P(Z_t|X_t^*, Y_t, Z_{t-1}) = P(Z_t|X_t^*)$.*
*(ii) For all $t$, the $K \times K$ matrix $\mathbf{G}_{Z_t|Z_{t-1}}$, with $(i,j)-th$ entry equal to $Pr(Z_t = i|Z_{t-1} = j)$, is invertible.*
*(iii) $E[Z_t|X_t^*]$ is increasing in $X_t^*$.*

As with Assumption 1, this assumption involves an important exclusion restriction that, conditional on $X_t^*$, the eye movement $Z_t$ in period $t$ is independent of $Z_{t-1}$. This serial independence assumption is, to some extent, imposed by construction in the experimental

setup, because we require subjects to "fix" their gaze in the middle of the computer screen at the beginning of each period. This should remove any inherent serial correlation in eye movements which is not related to the learning task.[17]

The invertibility assumption (3(i)) is made on the observed matrix $\mathbf{G}_{Z_t|Z_{t-1}}$ with elements equal to the conditional distribution of $Z_t|Z_{t-1}$; hence it is testable. Assumption 3(ii) "normalizes" the beliefs $X_t^*$ in the sense that, because large values of $Z_t$ imply that the subject gazed longer at blue, the monotonicity assumption implies that larger values of $X_t^*$ denote more "positive" beliefs that the current state is blue.

**Assumption 4** *The choice probabilities $P(Y_t|X_t^*)$, learning rules $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$, and measurement probabilities $P(Z_t|X_t^*)$ are the same for all subjects and trials $t$.*

This "stationarity" assumption justifies pooling the data across all subjects and trials for estimating the model. As with the other assumptions, it is motivated by the structure of optimal decision-making discussed in Section 2.2.1 above, where both the Bayesian belief-updating rule (Eq. (A.1) in Appendix A.1) and optimal choice rules in Figure 2.2 are indeed stationary.

### 2.3.1 Identification

In this section, we will use the shorthand notation $f(\cdots)$ to denote generically a probability distribution. For identification, we exploit the following relationship: conditional on $(R_{t-1})$, we have

$$f(Y_t, Z_t, X_t^*|Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) = f(Y_t, Z_t, X_t^*|Y_{t-1}, R_{t-1}, X_{t-1}^*). \tag{2.4}$$

[17] At the same time, we have also estimated models in which we allow $Z_t$ and $Z_{t-1}$ to be correlated, even conditional on $X_t^*$. These are reported in Appendix D.4. The results there show that the results are quite similar, for different values of $Z_{t-1}$, which imply that Assumption 3 is quite reasonable.

Abusing terminology somewhat, we call this a "first-order Markov" property, because the model exhibits only a one-period history dependence:

$$
\begin{aligned}
&f(Y_t, Z_t, X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \\
=&f(Y_t | Z_t, X_t^*, Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \cdot f(Z_t | X_t^*, Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \cdot f(X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \\
=&f(Y_t | X_t^*) \cdot f(Z_t | X_t^*) \cdot f(X_t^* | X_{t-1}^*, R_{t-1}, Y_{t-1}) \\
=&f(Y_t, Z_t, X_t^* | Y_{t-1}, R_{t-1}, X_{t-1}^*).
\end{aligned}
$$

In the above, the second equality applies Assumptions 1, 2, and 3.

The unknown functions we want to identify and estimate are:
(i) $f(Y_t|X_t^*)$, the *choice probabilities*;
(ii) the *learning rule* $f(X_t^*|X_{t-1}^*, Y_{t-1}, R_{t-1})$; and
(iii) the *measurement probabilities* $f(Z_t|X_t^*)$, the mapping between the auxiliary measure $Z_t$ and the unobserved beliefs $X_t^*$.

The nonparametric identification of these elements follows from an application of results from Hu (2008), and follows two main steps. Before presenting it, we note that, despite its simplicity, this model is not straightforward to estimate: given data on subjects' choices and rewards, we need to estimate choice probabilities conditional on subjects' beliefs, even though these beliefs are not only unobserved, but also changing over time.

**Step one: identification of choice probabilities $\mathbf{P}(\mathbf{Y_t}|\mathbf{X_t^*})$ and measurement probabilities $\mathbf{P}(\mathbf{Z_t}|\mathbf{X_t^*})$.** Consider the joint density $f(Z_t, Y_t|Z_{t-1})$, which is solely a function

of variables observed in the data. We can factor this density as follows:

$$\begin{aligned} f(Z_t, Y_t | Z_{t-1}) &= \sum_{X_t^*} f(Z_t, Y_t, X_t^* | Z_{t-1}) \\ &= \sum_{X_t^*} f(Z_t | Y_t, X_t^*, Z_{t-1}) f(Y_t, X_t^* | Z_{t-1}) \\ &= \sum_{X_t^*} f(Z_t | Y_t, X_t^*, Z_{t-1}) f(Y_t | X_t^*, Z_{t-1}) f(X_t^* | Z_{t-1}) \\ &= \sum_{X_t^*} f(Z_t | X_t^*) f(Y_t | X_t^*) f(X_t^* | Z_{t-1}) \end{aligned}$$

where the last equality applies to Assumptions 1 and 3.

For any fixed $Y_t = y$, then, we can write the above in matrix notation as:

$$\mathbf{A}_{y, Z_t | Z_{t-1}} = \mathbf{B}_{Z_t | X_t^*} \mathbf{D}_{y | X_t^*} \mathbf{C}_{X_t^* | Z_{t-1}}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all $K \times K$ matrices, and $\mathbf{D}$ is a $K \times K$ diagonal matrix. These are defined as:

$$\begin{aligned} \mathbf{A}_{y, Z_t | Z_{t-1}} &= \left[ f_{Y_t, Z_t | Z_{t-1}}(y, i | j) \right]_{i,j} \\ \mathbf{B}_{Z_t | X_t^*} &= \left[ f_{Z_t | X_t^*}(i | k) \right]_{i,k} \\ \mathbf{C}_{X_t^* | Z_{t-1}} &= \left[ f_{X_t^* | Z_{t-1}}(k | j) \right]_{k,j} \\ \mathbf{D}_{y | X_t^*} &= \begin{bmatrix} f_{Y_t | X_t^*}(y|1) & 0 & 0 \\ 0 & f_{Y_t | X_t^*}(y|2) & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & f_{Y_t | X_t^*}(y|K) \end{bmatrix} \end{aligned} \tag{2.5}$$

Similarly to the above, we can derive that

$$\mathbf{G}_{Z_t | Z_{t-1}} = \mathbf{B}_{Z_t | X_t^*} \mathbf{C}_{X_t^* | Z_{t-1}}$$

where $\mathbf{G}$ is likewise a $K \times K$ matrix, defined as

$$\mathbf{G}_{Z_t|Z_{t-1}} = \left[f_{Z_t|Z_{t-1}}(i|j)\right]_{i,j}. \tag{2.6}$$

From Assumption 3(i), we combine the two previous matrix equalities to obtain

$$\mathbf{A}_{y,Z_t|Z_{t-1}}\mathbf{G}^{-1}_{Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*}\mathbf{D}_{y|X_t^*}\mathbf{B}^{-1}_{Z_t|X_t^*}. \tag{2.7}$$

This is an eigenvalue decomposition of the matrix $\mathbf{A}_{y,Z_t|Z_{t-1}}\mathbf{G}^{-1}_{Z_t|Z_{t-1}}$, which can be computed from the observed data sequence $\{Y_t, Z_t\}$.[18] This shows that from the observed data, we can identify the matrices $\mathbf{B}_{Z_t|X_t^*}$ and $\mathbf{D}_{y|X_t^*}$, which are the matrices with entries equal to (respectively) the measurement probabilities $P(Z_t|X_t^*)$ and choice probabilities $P(Y_t|X_t^*)$.

In order for this identification argument to be valid, the eigendecomposition in Eq. (2.7) must be unique. This requires the eigenvalues in this decomposition (corresponding to choice probabilities $P(y|X_t^*)$) to be distinctive; that is, $P(y|X_t^*)$ should vary in $X_t^*$. This is ensured by Assumption 1. Furthermore, even if the eigendecomposition is unique, the representation in Eq. (2.7) is invariant to the ordering (or permutation) and scalar normalization of eigenvectors. Assumption 3(ii) imposes the correct ordering on the eigenvectors: specifically, it implies that columns with higher average value correspond to larger value of $X_t^*$. Finally, because the eigenvectors in the decomposition correspond to the conditional probabilities $P(Z_t|X_t^*)$, it is appropriate to normalize each column so that it sums to one. Hence, the uniqueness of the eigendecomposition, coupled with the ordering and normalization assumptions, ensure that the choice probabilities, measurement probabilities, and learning rules can be uniquely identified from the observed matrices $\mathbf{A}$ and $\mathbf{G}$.

[18] Note that, from Eq. (2.6), the invertibility of $\mathbf{G}$ (which is Assumption 3(i)) implies the invertibility of $\mathbf{B}$.

**Step two: identification of learning rule probabilities $\mathbf{P(X^*_{t+1}|X^*_t, R_t, Y_t)}$.** Again, start with a factorization

$$\begin{aligned} f(Z_{t+1}, Y_t, R_t, Z_t) &= \sum_{X^*_t} \sum_{X^*_{t+1}} f(Z_{t+1}, X^*_{t+1}, Y_t, X^*_t, R_t, Z_t) \\ &= \sum_{X^*_t} \sum_{X^*_{t+1}} f(Z_{t+1}|X^*_{t+1}) f(X^*_{t+1}|Y_t, X^*_t, R_t) f(Z_t|X^*_t) f\left(Y_t, X^*_t, R_t\right) \\ &= \sum_{X^*_t} \sum_{X^*_{t+1}} f(Z_{t+1}|X^*_{t+1}) f(X^*_{t+1}, Y_t, X^*_t, R_t) f(Z_t|X^*_t) \end{aligned}$$

where the second equality applies to Assumptions 1, 2, and 3. Then, for any fixed $Y_t = y$ and $R_t = r$, we have the matrix equality

$$\mathbf{H}_{Z_{t+1},y,r,Z_t} = \mathbf{B}_{Z_{t+1}|X^*_{t+1}} \mathbf{L}_{X^*_{t+1},X^*_t,y,r} \mathbf{B'}_{Z_t|X^*_t}.$$

The matrices $\mathbf{H}$ and $\mathbf{L}$ are $K \times K$ matrices defined as

$$\begin{aligned} \mathbf{H}_{Z_{t+1},y,r,Z_t} &= \left[f_{Z_{t+1},Y_t,R_t,Z_t}(i,y,r,j)\right]_{i,j} \\ \mathbf{L}_{X^*_{t+1},X^*_t,y,r} &= \left[f_{X^*_{t+1},X^*_t,Y_t,R_t}(i,j,y,r)\right]_{i,j}. \end{aligned} \tag{2.8}$$

Assumption 4 ensures that $\mathbf{B}_{Z_{t+1}|X^*_{t+1}} = \mathbf{B}_{Z_t|X^*_t}$. Hence, we can obtain $\mathbf{L}_{X^*_{t+1},X^*_t,y,r}$ (corresponding to the learning rule probabilities) directly from

$$\mathbf{L}_{X^*_{t+1},X^*_t,y,r} = \mathbf{B}^{-1}_{Z_{t+1}|X^*_{t+1}} \mathbf{H}_{Z_{t+1},y,r,Z_t} [\mathbf{B'}_{Z_t|X^*_t}]^{-1}. \tag{2.9}$$

This result implies that two periods of data $(Z_t, Y_t, R_t), (Z_{t-1}, Y_{t-1}, R_{t-1})$ are sufficient to identify and estimate this learning model.

## 2.4 Estimation

Our estimation procedure mimics the two-step identification argument from the previous section. That is, for fixed values of $(y, r)$, we first form the matrices $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{H}$ (as defined previously) from the observed data, using sample frequencies to estimate the corresponding probabilities. Then we obtain the matrices $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{L}$ using the matrix manipulations in Eqs. (2.7) and (2.9).

To implement this, we assume that the eye-movement measures $Z_t$ and the unobserved beliefs $X_t^*$ are discrete, and take three values.[19] One technical feature is that, because all the elements in the matrices of interest $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{L}$ correspond to probabilities, they must take values within the unit interval. However, in the actual estimation, we found that occasionally the estimates do go outside this range. In these cases, we obtained the estimates by a least-squares fitting procedure, where we minimized the element-wise sum-of-squares corresponding to Eqs. (2.7) and (2.9), and explicitly restricted each element of the matrices to lie $\in [0, 1]$. This was not a frequent recourse; only a handful of the estimates reported below needed to be restricted in this manner.[20]

In addition, while the identification argument above was "cross-sectional" in nature, being based upon two observations of $\{Y_t, Z_t, R_t\}$ per subject, in the estimation we exploited the long time series data we have for each subject, and pooled every two time-contiguous observations $\{Y_{i,r,\tau}, Z_{i,r,\tau}, R_{i,r,\tau}\}_{\tau=t-1}^{\tau=t}$ across all subjects $i$, all blocks $r$, and all trials $\tau = 2, \ldots, 25$. Formally, this is justified under the assumption that the process $\{Y_t, Z_t, R_t\}$ is stationary and ergodic for each subject and each block; under these assumptions, the ergodic theorem ensures that the (across time and subjects) sample frequencies used to construct

[19] Since the eye-movement measure $\tilde{Z}_t$ is continuous, we must discretize it for estimation. We leave the details of our discretization procedure (including a discussion of whether three points of discretization is appropriate) in Appendix A.3.

[20] In principle, because we do not impose *a priori* that the estimated probabilities must lie in [0,1] in estimated, we could use these overidentifying restrictions to test the model. While this works as an informal "eyeball" test, developing the formal sampling theory behind such a test seems difficult, due to the complexities in characterizing the behavior of a test statistic at the boundary values of 0 and 1, and so we do not pursue it here.

the matrices **A**, **G**, and **H** converge towards population counterparts.[21]

### 2.4.1 Estimation results

Tables 2.3 and 2.4 present estimation results. Both $X_t^*$ and $Z_t$ are discretized to take values $\{1, 2, 3\}$. We interpret $X^* = 1, 3$ as indicative of "strong beliefs" favoring (respectively) green and blue, while the intermediate value $X^* = 2$ indicates that the subject is "not sure".[22]

Table 2.3 contains the estimates of the choice and measurement probabilities.[23] The first and last columns of the panels in this table indicate that choices and eyes movements are closely aligned with beliefs, when beliefs are sufficiently strong (i.e., are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the probability of choosing a color contrary to beliefs—which is called the "exploration probability" in the literature—is small, being equal to 1.3% when $X_t^* = 1$, and only 0.64% when $X_t^* = 3$.

When $X_t^* = 2$, however, suggesting that the subject is unsure of the state, there is a slight bias in choices towards "blue", with $Y_t = 2$ roughly 56% of the time. The bottom panel indicates that when subjects are not sure, they tend to split their gaze more evenly between the two colors (i.e., $Z_t = 2$) around 63% of the time.

The learning rule estimates are presented in Table 2.4. The left columns show how beliefs are updated when "exploitative" choices (i.e., choices made in accordance with beliefs) are

[21] Results from Monte Carlo simulations (available from the authors on request) show that the estimation procedure produces accurate estimates of the model components, with the differences between the estimated and actual values usually on the order of magnitude of $10^{-1}$ times the parameter value.

[22] We have tried to re-estimate the model allowing for more belief states ($\geq 4$), but the results we obtained was not encouraging. This is due to our relatively small sample size; since our estimation approach is nonparametric, it is difficult to obtain reliable estimates with modest sample sizes. At the same time, as we pointed out above, statistical evidence indicates that it is sufficient to discretize the eye-movement measure $Z_t$ into three values, which implies that beliefs $X_t^*$ should not take more than 3 values.

[23] We also considered a robustness check against the possibility that subjects' fixations immediately before making their choices coincide exactly with their choice. While this is not likely in our experimental setting, because subjects were required to indicate their choice by pressing a key on the keyboard, rather than clicking on the screen using a mouse, we nevertheless re-estimated the models but eliminating the last segment of the reaction time in computing the $Z_t$. The results are very similar to the reported results, both qualitatively and quantitatively.

Table 2.3: Estimates of choice and measurement probabilities

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Each column sums to one.

$P(Y_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| $Y_t = 1$ (green) | 0.9866 (0.0561) | 0.4421 (0.1274) | 0.0064 (0.0146) |
| 2 (blue) | 0.0134 | 0.5579 | 0.9936 |

$P(Z_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| $Z_t = 1$ (green) | 0.8639 (0.0468) | 0.2189 (0.1039) | 0.0599 (0.0218) |
| 2 (middle) | 0.0815 (0.0972) | 0.6311 (0.1410) | 0.0980 (0.0369) |
| 3 (blue) | 0.0546 (0.0581) | 0.1499 (0.1206) | 0.8421 (0.0529) |

Table 2.4: Estimates of learning (belief-updating) rules
Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Each column sums to one.

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =1(lose), $y$ =1(green)

| $X^*_t$ | 1(green) | 2 (not sure) | 3(blue) |
|---|---|---|---|
| $X^*_{t+1}=1$ (green) | 0.5724 (0.0694) | 0.3075 (0.0881) | 0.1779 (0.2257) |
| 2 (not sure) | 0.0000[a] (0.0662) | 0.3138 (0.1042) | 0.4002 (0.2284) |
| 3 (blue) | 0.4276 (0.0624) | 0.3787 (0.0945) | 0.4219 (0.2195) |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =2(win), $y$ =1(green)

| $X^*_t$ | 1(green) | 2 (not sure) | 3(blue) |
|---|---|---|---|
| $X^*_{t+1}=1$ (green) | 0.8889 (0.0894) | 0.6621 (0.1309) | 0.8242 (0.2734) |
| 2 (not sure) | 0.0000 (0.0911) | 0.2702 (0.1297) | 0.1758 (0.1981) |
| 3 (blue) | 0.1111 (0.0340) | 0.0678 (0.0485) | 0.0000 (0.1876) |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =1(lose), $y$ =2(blue)

| $X^*_t$ | 3(blue) | 2 (not sure) | 1(green) |
|---|---|---|---|
| $X^*_{t+1}=3$ (blue) | 0.5376 (0.0890) | 0.2297 (0.0731) | 0.2123 (0.1436) |
| 2 (not sure) | 0.0458 (0.0732) | 0.2096 (0.0958) | 0.1086 (0.1524) |
| 1 (green) | 0.4166 (0.0874) | 0.5607 (0.0968) | 0.6792 (0.1881) |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =2(win), $y$ =2(blue)

| $X^*_t$ | 3(blue) | 2 (not sure) | 1(green) |
|---|---|---|---|
| $X^*_{t+1}=3$ (blue) | 0.8845 (0.1000) | 0.6163 (0.1136) | 0.6319 (0.1647) |
| 2 (not sure) | 0.0000 (0.0968) | 0.3558 (0.1160) | 0.3566 (0.1637) |
| 1 (green) | 0.1155 (0.0499) | 0.0279 (0.0373) | 0.0116 (0.0679) |

[a]This estimate, as well as the other estimates in this table which are equal to zero, resulted from applying the constraint that probabilities must lie between 0 and 1. See the discussion in Section 4 for more details.

taken, and illustrate an important asymmetry in subjects' belief-updating rules. When current beliefs indicate "green" ($X_1^* = 1$) and green is chosen ($Y_t = 1$), beliefs evolve asymmetrically depending on the reward: if $R_t = 2$ (high reward), then beliefs update towards green with probability 89%; however, if $R_t = 1$ (low reward), then belief still stay at green with probability 57%. This tendency of subjects to update up after successes, but not update down after failures also holds after a choice of "blue" (as shown in the left-hand columns of the bottom two panels in Table 2.4): there, subjects update their belief on blue up to 88% following a success ($R_t = 2$), but still give the event blue a probability of 53% following a failure ($R_1 = 1$). This muted updating following failures is a distinctive feature of our learning rule estimates and, as we will see below, is at odds with optimal Bayesian belief updating.

The results in the right-most columns describe belief updating following "explorative" (contrarian to current beliefs) choices. For instance, considering the top two panels, when current beliefs are favorable to "blue" ($X_t^* = 3$), but "green" is chosen, beliefs update more towards "green" ($X_{t+1}^* = 1$) after a low rather than high reward (82% vs. 18%). However, the standard errors (computed by bootstrap) of the estimates here are much higher than the estimates in the left-hand columns; this is not surprising, as the choice probability estimates in Figure 2.3 show that explorative choices occur with very low probability, leading to imprecision in the estimates of belief-updating rules following such choices.

The second columns in these panels show how beliefs evolve following (almost-) random choices. Again considering the top two panels, we see that when current beliefs are unsure ($X_t^* = 2$), there is stronger updating towards "green" when green choice yielded the higher reward (66% vs. 31%). The results in the bottom two panels are very similar to those in the top two panels, but describe how subjects update beliefs following choices of "blue" ($Y_t = 2$).

## 2.5 How optimal are estimated learning rules?

In the remainder of the paper, we compare our estimated learning rules to alternative learning rules which have been considered in the literature. We consider four alternative parametric learning rules: (i) the *optimal dynamic Bayesian* model, which is the model discussed in Section 2.2.1 above; (ii) a *pseudo-Bayesian* model, which is a version of the optimal Bayesian model in which the decision rules are smoothed relative to the step-function decision rules in the optimal model (cf. Figure 2.2); (iii) *reinforcement learning* (cf. Sutton and Barto (1998)); and (iv) *win-stay*, a simple choice heuristic whereby subjects replay successful strategies. All of these models, except (i), contain unknown model parameters, which we estimated using the choice data from the experiments. Complete details on these models, and the estimated model parameters, are given in Appendix A.2.

The relative optimality of each learning model was assessed via simulation. For each model, we simulated 100,000 sequences (each containing eight blocks of choices, as in the experiments) of rewards and choices, and computed the distributions of payoffs obtained by agents. The empirical quantiles of these distributions are presented in Table 2.5.

Table 2.5: Simulated payoffs from learning models

| | Optimal Bayesian[a] | Nonparametric | Pseudo-Bayesian | Reinforcement Learning | Win-stay |
|---|---|---|---|---|---|
| 5-%tile | \$5 | \$1 | \$2 | \$1 | \$1 |
| 25-%tile | \$12 | \$8 | \$9 | \$8 | \$8 |
| 50-%tile | \$17 | \$13 | \$14 | \$13 | \$13 |
| 75-%tile | \$22 | \$18 | \$19 | \$18 | \$18 |
| 95-%tile | \$29 | \$25 | \$26 | \$25 | \$25 |

[a]As described in Section 2.2.1

Reinforcement-learning, pseudo-Bayesian, and win-stay models are described in Appendix A.2. For each model, the quantiles of the simulated payoff distribution (across 100,000 simulated choice/reward sequences) is reported.

As we expect, the optimal Bayesian model generates the most revenue for subjects; the simulated payoff distribution for this model stochastically dominates the other models, and the median payoff is \$17. The other models perform almost identically, with a median

payoff around \$3-\$4 less than the Bayesian model (or about two cents per choice). This difference accounts for about 25% of typical experimental earnings (not counting the fixed show-up fee).

In the next section, we look for explanations for the differences (and similarities) in performance among the alternative learning models by comparing the belief-updating and choice rules across the different models.

### 2.5.1 Comparing choice and belief-updating rules across different learning models

For the optimal Bayesian and reinforcement-learning models, we can recover the "beliefs" corresponding to the observed choices and rewards, and compare them to the beliefs from the nonparametric learning model.[24] Appendix A.2 contains additional details on how the beliefs were derived for the learning models.

In Table 2.6, we present some summary statistics for the implied beliefs from our nonparametric learning model (denoted $X_t^*$), vs. the Bayesian beliefs $B^*$ and the valuations $V^*$ in the reinforcement-learning model. For simplicity, we will abuse terminology somewhat and refer in what follows to $X^*$, $V^*$, and $B^*$ as the "beliefs" implied by, respectively, our nonparametric model, the reinforcement-learning model, and the Bayesian model.[25] This table contains eight panels.

Panel 1 gives the total tally, across all subjects, blocks, and trials, of the number of times the nonparametric beliefs $X^*$ took each of the three values. Subjects' beliefs tended to favor green and blue roughly equally, with "not sure" lagging far behind. The close split

[24] There are no beliefs in the win-stay model, which is a simple choice heuristic. The pseudo-Bayesian model has the same beliefs as the optimal Bayesian model (with the difference that the choice rule is smoothed).

[25] As we clarify in Appendix A.2, the nonparametric beliefs $X_t^*$ were estimated from a maximum likelihood procedure which ignores the implied correlation between choices and beliefs; this is because, given the estimates of the choice probabilities in Table 3, which showed that $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$, estimating beliefs $X_t^*$ based on observed choices $Y_t$ would lead to estimates of beliefs which practically coincide with choices (i.e., $X_t^* = Y_t$), an artificially good "fit" which we felt does not accurately represent the belief process of the subjects.

Table 2.6: Summary statistics for beliefs in three learning models

$X^*$: Beliefs from nonparametric model
$B^*$: Beliefs from Bayesian model
$V^*$: "Beliefs" (valuations) from reinforcement learning model **Panel 1**:

| $X^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| | 1878 (45%) | 366 (10%) | 1956 (45%) |

**Panel 2**:

| | mean | median | std. | 33%-tile | 33%-tile |
|---|---|---|---|---|---|
| $B^*$ (Bayesian Belief) | 0.4960 | 0.5000 | 0.1433 | 0.4201 | 0.5644 |
| $V^*(= V_b - V_g)$ | -0.0104 | 0 | 0.4037 | -0.2095 | 0.1694 |

See Appendix A.2 for details on computation of beliefs in these three learning models.

between "green" and "blue" beliefs is consistent with the notion that subjects have rational expectations, with flat priors on the unobserved state $S_1$ at the beginning of each block. The second panel shows analogous statistics for the beliefs from the reinforcement-learning and Bayesian models. The reinforcement-learning valuation measure $V^*$ appears largely symmetric and centered around zero, while the average Bayesian $B^*$ lies also around 0.5. Thus, on the whole, all three measures of beliefs appear equally distributed between "green" and "blue".

Next, we compare the learning rules from the nonparametric, (optimal) Bayesian, and reinforcement learning models. In order to do this, we discretized the beliefs in each model into three values, in proportions identical to the frequency of the different values of $X_t^*$ as reported in Table 2.6, and present the implied learning rules for each model.[26] These are shown in Table 2.7.

Comparing the three sets of learning rules, we see that the most striking difference between

[26]Specifically, we discretized the Bayesian (resp. reinforcement-learning) beliefs so that 45% of the beliefs fell in the $B_t^* = 1$(resp. $V_t^* = 1$) and $B_{t+1}^* = 3$(resp. $V_t^* = 3$) categories, while 10% fell in the intermediate $B_t^* = 2(X_t^* = 2)$ category, the same as for the nonparametric beliefs $X_t^*$ (cf. Panel 1 of Table 6). The results are even more striking when we discretized the Bayesian and reinforcement-learning beliefs so that 33% fell into each of the three categories.

Table 2.7: Learning (belief-updating) rules for alternative learning models

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =1(lose), $y$ =1(green)

| | **Optimal Bayesian Learning** | | | **Reinforcement Learning** | | |
|---|---|---|---|---|---|---|
| Beliefs $B^*_{t+1}$, $V^*_{t+1}$: | 1(green) | 2 (not sure) | 3(blue) | 1(green) | 2 (not sure) | 3(blue) |
| 1 (green) | 0.2878 | 0 | 0 | 0.6538 | 0 | 0 |
| 2 (not sure) | 0.1730 | 0 | 0 | 0.1381 | 0.0115 | 0 |
| 3 (blue) | 0.5392 | 1.0000 | 1.0000 | 0.2080 | 0.9885 | 1.0000 |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =2(win), $y$ =1(green)

| | **Optimal Bayesian Learning** | | | **Reinforcement Learning** | | |
|---|---|---|---|---|---|---|
| Beliefs $B^*_{t+1}$, $V^*_{t+1}$: | 1(green) | 2 (not sure) | 3(blue) | 1(green) | 2 (not sure) | 3(blue) |
| 1 (green) | 1.0000 | 1.0000 | 0.6734 | 1.0000 | 0.8818 | 0.6652 |
| 2 (not sure) | 0 | 0 | 0.1250 | 0 | 0.1182 | 0.1674 |
| 3 (blue) | 0 | 0 | 0.2016 | 0 | 0 | 0.1674 |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =1(lose), $y$ =2(blue)

| | **Optimal Bayesian Learning** | | | **Reinforcement Learning** | | |
|---|---|---|---|---|---|---|
| Beliefs $B^*_{t+1}$, $V^*_{t+1}$: | 3(blue) | 2 (not sure) | 1(green) | 3(blue) | 2 (not sure) | 1(green) |
| 3 (blue) | 0.3060 | 0 | 0 | 0.6576 | 0 | 0 |
| 2 (not sure) | 0.1601 | 0 | 0 | 0.1261 | 0.0109 | 0 |
| 1 (green) | 0.5338 | 1.0000 | 1.0000 | 0.2164 | 0.9891 | 1.0000 |

$P(X^*_{t+1}|X^*_t, y, r)$, $r$ =2(win), $y$ =2(blue)

| | **Optimal Bayesian Learning** | | | **Reinforcement Learning** | | |
|---|---|---|---|---|---|---|
| Beliefs $B^*_{t+1}$, $V^*_{t+1}$: | 3(blue) | 2 (not sure) | 1(green) | 3(blue) | 2 (not sure) | 1(green) |
| 3 (blue) | 1.0000 | 1.0000 | 0.6760 | 1.0000 | 0.8898 | 0.6983 |
| 2 (not sure) | 0 | 0.0000 | 0.1440 | 0 | 0.1102 | 0.1379 |
| 1 (green) | 0 | 0 | 0.1800 | 0 | 0 | 0.1638 |

them is in how beliefs update following unsuccessful choices (i.e., choices which yielded a negative reward). Comparing the Bayesian and the nonparametric learning rules (in Table 5), we see that Bayesian beliefs exhibit less "stickiness", or serial correlation, following unsuccessful choices. For example, consider the case of $(Y_t = 1, R_t = 1)$, so that an unsuccessful choice of green occurred in the previous period. The nonparametric learning rules (Table 5) show that the weight of beliefs remain on "green" $(X^*_{t+1} = 1)$ with 57% probability, whereas the Bayesian beliefs place only 28% weight on green. A similar pattern exists after an unsuccessful choice of blue, as shown in the left-hand column of the third panel.

On the other hand, the learning rules for the reinforcement-learning model (also reported in Table 2.7) are more similar to the nonparametric learning rule, especially following unsuccessful choices. Again, looking at the top panel, we see that following an unsuccessful choice of "green" $(Y_t = 1)$, subjects valuations are still favorable to green with probability 65%; this is comparable in magnitude to the 57% from the nonparametric learning rule. Similarly, after an unsuccessful choice of blue (third panel), valuations in the reinforcement-learning model still favor blue with probability 66%, again comparable to the 54% for the nonparametric model. It appears that the updating rules from the reinforcement-learning and nonparametric model share a common defect: a reluctance to "update down" following unsuccessful choices; this common defect relative to the optimal Bayesian model may explain the lower revenue generated by these models.

In Table 2.8 we compare the choice rules across the different models. As in the previous table, we discretized the beliefs from each model into three values. Comparing the top two panels, we see that, even though the belief-updating rule is the same for the optimal Bayesian and pseudo-Bayesian models, the choice rules are strikingly different. Evaluated at the estimated model parameter (discussed in Appendix A.2), choice probabilities in the pseudo-Bayesian model are practically invariant to the beliefs, and equal to around 50% for all values of beliefs.

In contrast, choice rules in the optimal Bayesian model are deterministic functions of beliefs.

Table 2.8: Choice rules for alternative learning models

| **Optimal Bayesian Learning** | | | |
|---|---|---|---|
| Beliefs $B_t^*$: | 1(green) | 2(not sure) | 3(blue) |
| $Y_t = 1$ (green) | 1.0000 | 0.5000 | 0.0000 |
| 2 (blue) | 0.0000 | 0.5000 | 1.0000 |

| **Pseudo-Bayesian Learning** | | | |
|---|---|---|---|
| Beliefs $B_t^*$: | 1(green) | 2(not sure) | 3(blue) |
| $Y_t = 1$ (green) | 0.5141 | 0.4996 | 0.4850 |
| 2 (blue) | 0.4859 | 0.5005 | 0.5150 |

| **Reinforcement Learning** | | | |
|---|---|---|---|
| Beliefs $V_t^*$: | 1(green) | 2(not sure) | 3(blue) |
| $Y_t = 1$ (green) | 0.7629 | 0.4939 | 0.2250 |
| 2 (blue) | 0.2371 | 0.5061 | 0.7750 |

Overall, the estimated choice rules in Table 2.3 are much closer to the optimal Bayesian model, than the pseudo-Bayesian model. This suggests that the lower payoffs from the estimated model relative to the optimal Bayesian model arise primarily not from the choice rules (which are very similar in the two models), but rather from the belief-updating rules (which are quite different, as discussed previously).

The bottom panel of Table 2.8 contains the choice rules for the reinforcement-learning model. As shown there, the choice rules are much smoother than in the optimal Bayesian model and the estimated model, but not as smooth as the pseudo-Bayesian model. This suggests that the similarities of the payoffs from the estimated model relative to reinforcement-learning (as shown in Table 2.5) arise mainly from the similarities in belief-updating rules, and less from the choice rules, which are quite different in the two models.

Finally, the similarity in payoffs between the nonparametric and win-stay models is not surprising because, as we showed in Section 2.2.3 above, the reduced-form choice behavior from the experimental data is in line with a "win-stay/lose-randomize" rule of thumb. Such behavior is confirmed in the formal parameter estimates for the win-stay model (presented in Appendix A.2.5) which show that, after receiving a positive reward, subjects tend to

repeat the previous choice with probability 87% while, after a negative reward, subjects essentially randomize. This asymmetry in choices following good/bad rewards echoes the nonparametric learning rules from Table 5, which showed that subjects "update down" much less following bad rewards than they "update up" following good rewards.

### 2.5.2 Are eye movements noisy measure of beliefs?

The empirical exercise we undertake in this paper hinges crucially on the assumption that eye movements are (noisy) measurements of the unobserved beliefs, and are not merely noisy measurements of choices. Having used the choice data to estimate beliefs for the nonparametric learning model, as well as the benchmark Bayesian model, we conclude the paper by using these beliefs to perform an assessment of this critical assumption.[27]

Independently of our empirical model and its underlying assumptions, eye movements are really related to some intuitive notion of how beliefs evolve? Using Bayesian beliefs as such an intuitive (and objective) measure of beliefs, we compare each Bayesian belief $B^*_{it}$ to the corresponding (undiscretized) eye movement measure $\tilde{Z}_{it}$ (as defined in Eq. (2.3)) recorded for that subject and trial. The graphs are presented in Figure 2.3. In the top graph, we see that $Z$ is clearly increasing with $B^*$, suggesting that eye movements track well a standard notion of beliefs.

Of course, this positive relationship could be spurious; if eye movements were not a noisy measure of beliefs, but rather of choice, then the graph here may be picking up simply the common dependence of both $\tilde{Z}$ and $B^*$ on choices. To address this, we consider, in the remaining two graphs in Figure 2.3, a plot of $(\tilde{Z}_t, B^*_t)$ values *conditional on the choice* $Y_t$; by conditioning on choice, we eliminate any variation in eye movements due to differences in choice.

For the most part, we see that the positive relationship between $\tilde{Z}_t$ and $B^*_t$ remains, even

[27] See also Appendix A.5 for another approach to assessing this assumption.

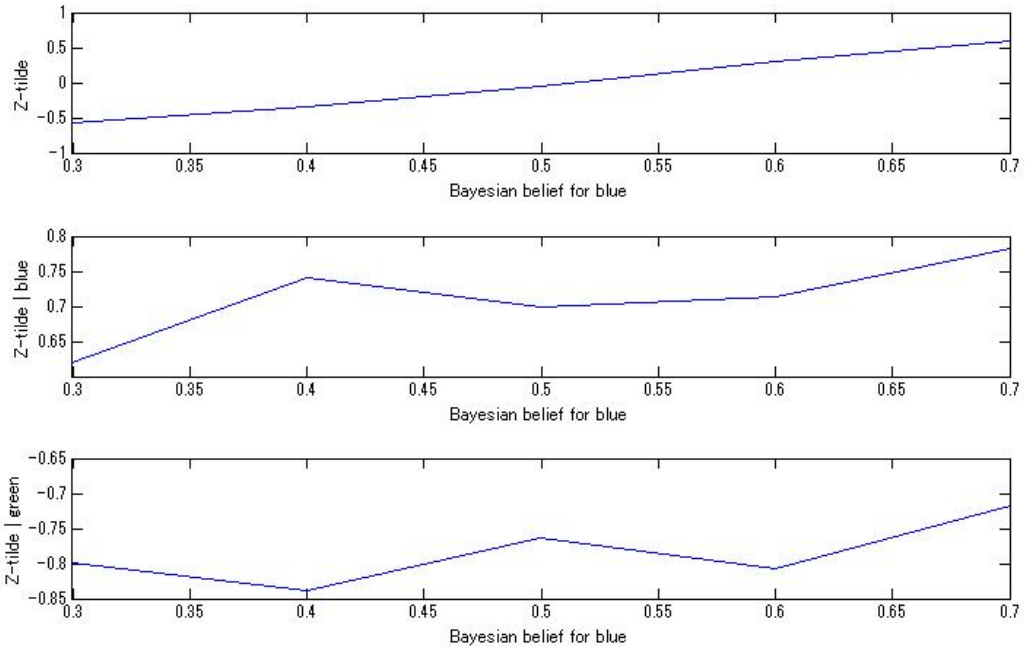



Figure 2.3: How eye movements track Bayesian beliefs

after conditioning on choice.[28] For example, the overall positive trend in the second graph suggests, reassuringly, that the eye tracking measure $\tilde{Z}_t$ more strongly favored blue (i.e., $\tilde{Z}_t$ takes large values) when there was strong evidence that blue was the good arm (i.e., $B_t^*$ is large), than when there is only weaker evidence (i.e., $B_t^*$ is small), even after controlling for the relationship between eye movements and choices. This is solid evidence that eye movements are related to some intuitive notion of how beliefs behave, and are not simply noisy measures of choices.[29]

[28] Moreover, the data will be less concentrated in the region of small values of $B^*$ in the second graph (when blue would tend not to be chosen), and large values of $B^*$ in the bottom graph (where green would tend not to be chosen). This may explain the kinks in the graphs in those regions.

[29] See also Appendix A.5 for an alternative assessment of this.

## 2.6 Conclusions

In this paper, we estimate learning rules nonparametrically from data drawn from experiments of multi-armed bandit problems. The experimental data are augmented by measurements of subjects' eye movements from an eye-tracker machine, which play the role of auxiliary measures of subjects' beliefs. Our estimated learning rules have some distinctive features—notably that subjects tend to update asymmetrically after unsuccessful choices as compared to successful choices. The profits from following the estimated learning and decision rules are smaller than what would be obtained from an optimal Bayesian learning model (about $4 less for each subject, at the median), and comparable to the profits obtained from three other parametric models: a reinforcement-learning model, a "pseudo"-Bayesian model, and a win-stay choice heuristic. Relative to the optimal Bayesian model, the belief-updating rules from the nonparametric and reinforcement-learning model share a common feature that subjects appear reluctant to "update down" following unsuccessful choices; this may explain the sub-optimality of these models (in terms of profits).

Our nonparametric estimator for subjects' choice probabilities and learning rules is easy to implement, involving only elementary matrix operations. Furthermore, from a methodological point of view, the *modus operandi* used in this paper—the nonparametric estimation of learning models using experimental data—appears to be a portable idea which can be potentially applied more broadly to other experiments involving dynamic decision problems.

# Chapter 3

# Learning Under Outlier Risk in a Non-Gaussian World

## 3.1 Introduction

Predicting price dynamics, especially after extremely large price changes such as bubble bursts, has been one of the central questions in financial economics, hence intensive studies has been done, often referred as heavy-tailed returns or outlier risk problem (Taleb (2008)). However, it is still elusive how and what kind of dynamics people see behind extremely large price changes, either some systemic changes or just an observation error without little structural change. In traditional financial economics, a large stock price change occurs as a result of its fundamental change (Cochrane (2005)), which derives that the price follows unpredictable random walk process (Malkiel (1999)). On the contrary in behavioral finance, a large price deviation could occur due to a temporal overreaction without any fundamental change (De Bondt and Thaler (1989), Bondt and Thaler (1985), Shiller (2003)) and the deviation would disappear with a predictable reversion to its fundamental value sooner or later.

Recent studies in decision neuroscience (Yu and Dayan (2005), Aston-Jones and Cohen (2005), Sara (2009)) and psychology (Kruschke and Johansen (1999), Griffin and Tversky

(1992), Gilovich, Vallone, and Tversky (1985)) revealed that salient information changes often alter mental models in the human mind (Gentner and Stevens (1983), Craik (1943)), suggesting that humans are more likely to see fundamental changes behind salient, large price changes; In natural environment, salient information generally have been critical signals which warn animals against "fundamental" changes of the surrounding environment which could be a thread of their lives, hence humans, or animals in general, might have developed a cognitive ability to inevitably pay attention to salient information in the long evolutionary histories (Bromberg-Martin, Matsumoto, and Hikosaka (2010)).

However as behavioral finance argues, in the modern financial market stock prices often show salient outliers which are seemingly unrelated to fundamental changes, and reverting to the original price level either in the short term (Bremer and Sweeney (1991)) or in the long term (the 2009 correction to the 2008 downfall in average world stock prices being one example). I speculate that the counterintuitive contradiction between the modern financial markets and natural environment could be a potential source for suboptimal behavior of humans in price predictions in the financial markets; Human might tend to see fundamental changes behind outliers even without them in financial markets, as similarily as animals see fundamental changes in environment from salient information.

In this study, I develop the optimal learning models for predicting price dynamics, or more precisely belief-updating process on price dynamics under the two kinds of outlier risk, as Bayesian reinforcement-learning models, solving them numerically using sequential Monte Carlo (SMC) sampling. More precisely, the two kinds of outlier risk considered here are followings;

1. "Fundamental" case in which, when the stock price shows a large movement, the subsequent change will tend to be around the new price level, depicting that the "fundamental" value of the stock has changed with the big leap.

2. "Anomaly" case in which, when the price shows a large movement, the price will tend to revert to the original level in the subsequent period; That is, the price deviates

from fundamental value temporally but revert to it, depicting a price "anomaly" in behavioral finance.

The modern financial markets have attracted a wide range of people, and as a result many people, who are not necessarily specialized in financial tradings, are involved in predicting price dynamics to earn their money. They are unavoidably faced with large price changes which are unclear to be whether permanent, or just temporal. It is not clear whether and how humans dissociate between these two types of outliers. Exploring the human nature of predicting price dynamics associated with large price changes is important and necessary for better decision-making for any financial participants, as its knowledge improves their decision-making as "nudge" (Thaler and Sunstein (2008)). It is also beneficial for policy makers to establish better market mechanisms such as circuit breaker, elaborated monetary policies and effective market interventions in a turbulent market, in order to maintain stable and sound financial markets.

In the next section, I present experimental evidence showing that humans are actually capable of discriminating the two price processes involving outliers with/without fundamental changes, which motivates our optimal learning model under outlier risk. In Section 3, I characterize the optimal learning models as Bayesian reinforcement learning models using non-Gaussian Kalman filter, and obtain optimal estimates of the hidden, fundamental value with sequential Monte Carlo (SMC) sampling. Then, I conclude in Section 4.

## 3.2 Human perception of outliers

Do people correctly attribute the sources of outliers, either fundamental changes or just observation noises as anomalies? To assess this issue, Bossaerts' research group conducts a behavioral experiment on human perception of the two kinds of outliers with/without fundamental changes[1]. They asked their subjects to predict movements of a Target depicting

[1]I acknowledge Peter Bossaerts and his research group for allowing me to use their data set.

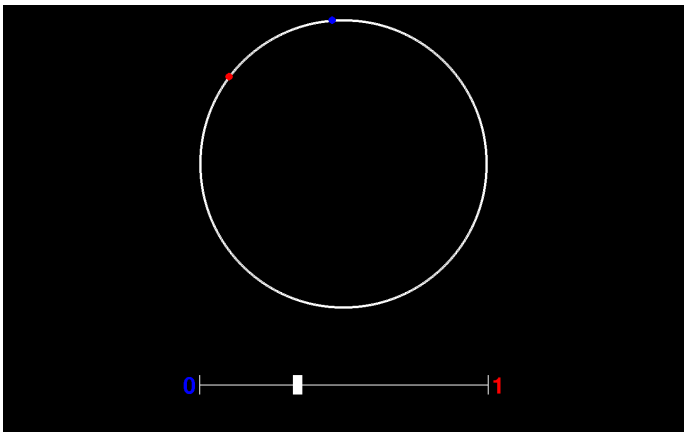



Figure 3.1: Target position prediction task

price dynamics including the two different kinds of outlier risk. In their Target position prediction task, subjects see a Target (red dot) that moves along a circle (Fig. 3.1), in both clockwise and counterclockwise direction. On the same circle, subjects will also observe a Robot (blue dot). The task will be to control the Robot by changing how much it responds to Target movements, using a slider underneath the circle (Fig. 3.1). Subjects' task is to forecast the Target's subsequent position and navigate the Robot to the predicted position. Moving the cursor towards "1" increases the adjustment of the Robot towards the Target's last move, while moving the cursor towards "0" decreases the Robot's adjustment. The experiment has two separated blocks, each having a distinctive experimental treatment for Target movement[2]:

1. "Fundamental" treatment in which, when the Target shows a large movement, the Target's subsequent position will tend to be around the new position, depicting that the "fundamental" value of the stock has changed with the big leap.

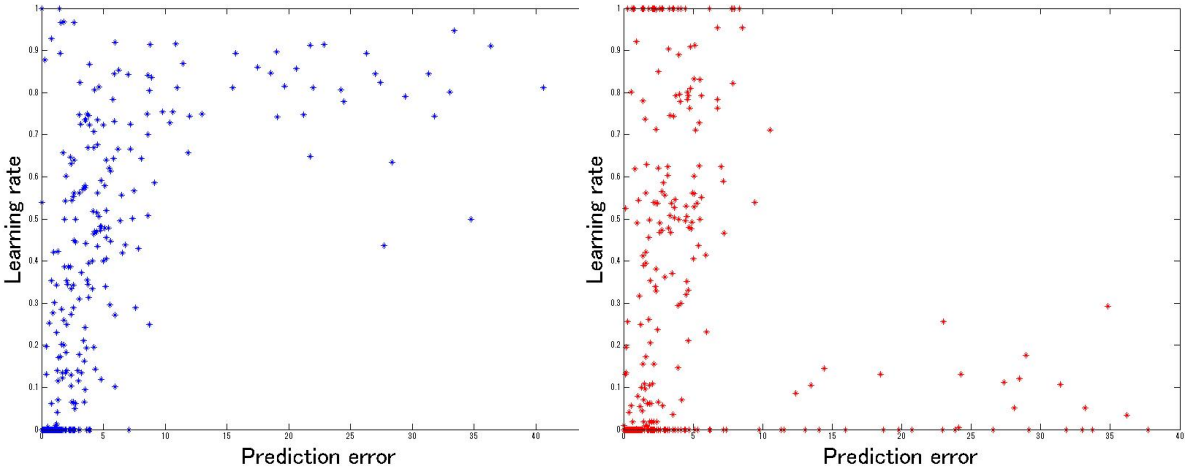
2. "Anomaly" treatment in which, when the price shows a large movement, the Target will tend to revert to the original level in the subsequent period; That is, the price deviates from fundamental value temporally but revert to it, depicting a price

[2]The full description of the two stochastic processes is in the Appendix B.1.

"anomaly" in behavioral finance.

Figure 3.2: Learning rate and prediction error

(Subject 8) Left: Fundamental treatment; Right: Anomaly treatment




x-axis: Learning rate; y-axis: Absolute value of prediction error

I plot learning rates and absolute values of prediction error for a subject (subject 9) in Figure 3.2. Here learning rate is the adjustment of the Robot, which controls how much subjects respond to the Target's last movement, whereas prediction error is the difference between Robot's current position and Target's new position, indicating how much her prior forecast was missed[3].

Apparently, the subject differentiates her forecasting strategy depending on the treatments as the distributions of learning rate are different according to the treatments with larger prediction error, suggesting that the subject certainly distinguishes the two stochastic processes. Although she takes a seemingly identical, random strategy for forecasting in both treatments when the prediction error is small (less than ten), faced with unexpected large movement of the Target she shows significantly larger learning rates in the fundamental treatment (left panel) than those in the anomaly treatment (right panel). The two plots

[3]The mathematical definition of learning rate and prediction error is fully discussed in the next section.

are clearly separable in the area of larger prediction error. Observing an outlier in Targets movement in the fundamental treatment, she tracks it a lot, expecting that the Target's subsequent position will be around its new position, which suggests that possibly she realizes the "fundamental" change of the Target's position behind the outliers. Contrarily in the anomaly treatment, she does not respond to outliers, implying that she expects their reversions to the previous positions in the subsequent trials, as she assumes that the outliers are just temporal "anomalies".

Not only the subject 9, but also most of the subjects more or less correctly distinguish the two underlying stochastic process successfully. Table 3.1 contains the evidence suggesting that most people more or less differentiate their predictions systematically in accordance with the treatments. The top row in the table contains the mean learning rate in the trials in which they are faced with large Target movements or jumps, conditional on the treatments either the fundamental or the anomaly treatment. The threshold value for jump detection is defined as the value three times larger than 3.04, where the value "3.04" is 84.13 percentile of the distribution of the prediction error. Three times larger than the 84.13 percentile is 3 in a standard Gaussian distribution (99.9 percentile), which is generally regard as outliers[4].

Comparing the first and second entries in this row, we see that the mean learning rate in the fundamental treatment shows much larger value (0.7192) than that in the anomaly treatment (0.3650). Thus, on average, people see permanent, "fundamental" changes of the Target's movements with outliers in the fundamental treatment while people see less in the anomaly treatment. In order to examine the difference of their forecasting strategies under the two treatments more statistically, I compare the distributions of learning rate in jump trials conditional on each treatment using Kolmogorov-Smirnov test. The last entry in Table 3.1 contains the p-value of Kolmogorov-Smirnov one-sided test that the cumulative distribution function of learning rate in jump trials in the anomaly treatment, denoted by

[4]The definition of outliers is consistent with Dayan and Yu (2003) who employ a threshold value that is three times as large as the standard error in a Gaussian distribution for outliers due to regime switches in volatility of the underlying process.

Table 3.1: Mean learning rate at jump trials

| | Mean: Fundamental | Mean: Anomaly | Kolmogorov-Smirnov test, $P$ |
|---|---|---|---|
| **Across All Subjects:** | 0.7192 | 0.3650 | 0.0000 |
| | (0.0133) | (0.0158) | ** |
| | | **For each individual subject:** | |
| *Subject*1: | 0.8297 | 0.5825 | 0.0068 |
| | (0.0333) | (0.0665) | ** |
| *Subject*2: | 0.9992 | 0.1835 | 0.0000 |
| | (0.0006) | (0.0510) | ** |
| *Subject*3: | 0.8558 | 0.5825 | 0.0027 |
| | (0.0262) | (0.0525) | ** |
| *Subject*4: | 0.3822 | 0.3557 | 0.0595 |
| | (0.0361) | (0.0483) | |
| *Subject*5: | 0.7751 | 0.2150 | 0.0000 |
| | (0.0344) | (0.0442) | ** |
| *Subject*6: | 0.4432 | 0.2356 | 0.0003 |
| | (0.0364) | (0.0401) | ** |
| *Subject*7: | 0.7507 | 0.2703 | 0.0000 |
| | (0.0574) | (0.0640) | ** |
| *Subject*8: | 0.7900 | 0.0772 | 0.0000 |
| | (0.0612) | (0.0297) | ** |
| *Subject*9: | 0.7920 | 0.0897 | 0.0000 |
| | (0.0181) | (0.0271) | ** |
| *Subject*10: | 0.8739 | 0.4705 | 0.0000 |
| | (0.0480) | (0.0429) | ** |
| *Subject*11: | 0.8507 | 0.8513 | 0.1732 |
| | (0.0000) | (0.0002) | |
| *Subject*12: | 0.7712 | 0.2503 | 0.0000 |
| | (0.0435) | (0.0661) | ** |
| *Subject*13: | 0.6650 | 0.3172 | 0.0000 |
| | (0.0619) | (0.0473) | ** |
| *Subject*14: | 0.9578 | 0.2856 | 0.0000 |
| | (0.0157) | (0.0652) | ** |
| *Subject*15: | 0.7563 | 0.3195 | 0.0001 |
| | (0.0573) | (0.0869) | ** |

Note: standard errors in parentheses

Asterisks show the rejection of the null with 5% (*) or 1% (**) significance level.

$F(x)$, is larger than that in the fundamental treatment, denoted by $G(x)$[5]. The P-value is smaller than 0.0001, showing that the two distribution $F(x)$ and $G(x)$ is not identical[6], and moreover, the former is larger than the latter, implying that people correctly see more fundamental changes in the fundamental treatment than in the anomaly treatment.

In the reminder of the Table 3.1, I also present the same statistics, the mean learning rate for each treatment and Kolmogorov-Smirnov test, calculated for each subject individually. Although there is some degree of heterogeneity in subjects' strategies, most of the subjects show much larger mean learning rate in the fundamental treatment than that in the anomaly treatment, except for subjects 4 and 11 who show almost identical learning rate under both treatments. Indeed, the Kolmogorov-Smirnov test rejects the null hypothesis for the thirteen subjects out of fifteen, which offers a concrete evidence that humans are able to discriminate the two kind of outlier risks.

## 3.3 Optimal learning models under outlier risk

Having seen the humans' capability of distinguishing the two kind of outlier risks with and without fundamental changes, in this section I describe optimal learning model for price dynamics which involves outlier risk, and discuss the difference from an benchmark learning in a Gaussian world, using Kalman filter framework in order to illuminate the features. Kalman filter is an efficient, statistical estimator predicting the dynamic system of hidden Markov process using noisy signals of it, however it is also known as a Bayesian optimal sequential estimator on the dynamics of hidden variable (Harrison and Stevens (1971), Meinhold and Singpurwalla (1983)). Indeed, Kalman filter is often applied in finance for short-term forecast of the unobservable fundamental value of financial securities (Wells (1996)) where prices are a noisy measure of fundamental value, including noises which

[5]The null hypothesis is that the two distributions are identical.

[6]There is no opportunity of order effect over the treatments since the experimental order of the treatments is reversed between odd numbered subjects (starting with the fundamental treatment) and even numbered subjects (starting with the anomaly treatment).

various microstructures of financial markets generate. As a benchmark for the comparison, I start with a standard Kalman filter without outlier risk to describe its basic structure and model assumptions.

After considering a standard Kalman filter, using the common structure of standard Kalman filter but modifying its key assumptions, I describe optimal learning models under outlier risk in two ways; Optimal learning under a price process in which outliers occur as fundamental changes, and a price process in which outlier occur without fundamental changes rather as temporal observation noise. Two main streams of the academic finance today—traditional finance and behavioral finance—have their own distinctive views on the price changes; Traditional finance theories assume a price change is a result of its fundamental change, while price change might occur without its fundamental value change in behavioral finance. The price change in the latter case is referred to price anomalies. I present the optimal learning under outlier risk both in fundamentals and price anomalies in the following sections.

### 3.3.1 Benchmark: optimal learning in a Gaussian world (Kalman filter)

Let $\{y_1, y_2, ..., y_t\}$ be a sequence of a scalar variable which is directly observable to econometricians. The variable $y_t$ depends on an unobservable variable $x_t$, known as the state of the nature. Here in the financial application, the observable variable is a stock price, and the unobservable variable is its fundamental value. The goal of Kalman filter is to make inferences about the unobservable $x_t$ from the history of $y_t$ up to time t. The relationship between $y_t$ and $x_t$ is linear and is specified by the observational equation,

$$y_t = x_t + e_t, \tag{3.1}$$

where the observation noise, denoted by $e_t$ is assumed to be a Gaussian distribution with mean zero and variance $\sigma_o^2$, denoted as $N(0, \sigma_o^2)$.

The dynamics of the state variable is defined by the system equation,

$$x_t = x_{t-1} + u_t, \tag{3.2}$$

where the system innovation, denoted by $u_t$ is assumed to be a Gaussian distribution with mean zero and variance $\sigma_i^2$, denoted as $N(0, \sigma_i^2)$. Importantly, the observation noise $e_t$ and system innovation $u_t$ are assumed to be independent.

Let $\hat{x}_t$ be the best estimate at time t for the state of the nature, as it minimizes the mean square error of predictions. Kalman filter gives the analytical solution for $\hat{x}_t$.

$$\begin{aligned} \hat{x_t} &= \hat{x}_{t-1} + \frac{\frac{1}{\sigma_o^2}}{\frac{1}{V_{t-1}^2+\sigma_i^2} + \frac{1}{\sigma_o^2}}(y_t - \hat{x}_{t-1}) \\ {V_t}^2 &= \frac{1}{\frac{1}{V_{t-1}^2+\sigma_i^2} + \frac{1}{\sigma_o^2}} \end{aligned} \tag{3.3}$$

where $V_t$ is the variable which characterizes the uncertainty of the estimates on $x_t$. More precisely in Bayesian framework, the Bayesian posterior belief on the state of nature is described by a Gaussian distribution with mean $\hat{x}_t$ and variance $V_t^2$, denoted by $N(\hat{x}_t, V_t^2)$. The value $V_0$ takes zero if the true value of the state of nature is fully revealed at time 0.

Interestingly, the Kalman filter equation is transformed into a reinforcement learning framework, which is often referred as Bayesian reinforcement-learning. To transform the equations, I define a variable "prediction error", which is commonly employed in machine learning.

Here, prediction error, denoted by $\delta_t$, is the difference between the predicted value $\hat{x}_{t-1}$ at time t-1 and the observed value $y_t$ at time t.

$$\delta_t = y_t - \hat{x}_{t-1}, \tag{3.4}$$

As it is verified with a little computation below, the expected value of the prediction error at time t-1 (conditioned on the information set $I_{t-1}$ at time t-1) is zero, implying that the prediction error takes positive (negative) value when the realized value $y_t$ is greater (smaller) than the expected value.

$$\begin{aligned} E[\delta_t|I_{t-1}] &= E[y_t - \hat{x}_{t-1}|I_{t-1}] \\ &= E[x_t - \hat{x}_{t-1}|I_{t-1}] \\ &= E[x_{t-1} - \hat{x}_{t-1}|I_{t-1}] \\ &= 0. \end{aligned} \tag{3.5}$$

Now the first equation in Kalman filter (3.3) is transformed in a reinforcement-learning framework with prediction error $\delta_t$.

$$\hat{x_t} = \hat{x}_{t-1} + A_t \cdot \delta_t, \tag{3.6}$$

The first equation describes how the predicted value for $x_t$ is updated with the feedback of prediction error via Rescorla and Wagner rule (Rescorla and Wagner (1972))[7]. Here $A_t$ is a learning rate, or Kalman gain, which controls the magnitude of feedback on value update from prediction error.

[7]For technical details on reinforcement learning, see Sutton and Barto (1998).

$$A_t = \frac{\frac{1}{\sigma_o^2}}{\frac{1}{V_{t-1}^2+\sigma_i^2} + \frac{1}{\sigma_o^2}}. \tag{3.7}$$

The remarkable feature of the Kalman filter in a Gaussian world is that the learning rate $A_t$ is a constant function with respect to prediction error; Even if the prediction error is extremely large, the learning rate does not change at all. Rather than prediction error, the determinants of the learning rate in a Gaussian world are the uncertainty of the prior belief $\delta_t$, and the variances of the observation error $e_t$ and the system innovation $u_t$. Besides the contribution of the uncertainty of the prior belief $\delta_t$, the learning rate is determined by the relative strength of the two signals, namely the observation noise and system innovation. For example, larger variations in system innovation strengthens its relative contribution to the transition of $x_t$, hence optimally the larger part of the prediction error is attributed to the system innovation and vice versa.

As I discuss in the following sections, in the non-Gaussian, heavy-tailed world, the learning rate is no more just a constant function with respect to prediction error. Hereafter I compare the standard Kalman filter in a Gaussian world with those in a non-Gaussian world with outlier risk, especially focusing on the functional form of the optimal learning rate.

### 3.3.2 Optimal learning under outlier risk in fundamental value transition

Traditional finance theory assumes that price changes are caused by its fundamental value changes (Cochrane (2005), Malkiel (2003)). The outlier risk is included in the dynamics of fundamental value, and the outliers in price changes are for the most part due to its fundamental changes. In order to allow the occasional outliers in fundamental values in the Kalman filter framework, the assumption on the system innovation in the standard Kalman filter is modified to incorporate heavy-tailed distribution. As is often employed as a heavy-tailed distribution in finance (Behr and Pötter (2009)), here I use a finite-mixture

of Gaussian distributions model, which is first introduced as models of return distribution by Kon (2012). Kon (2012) examines daily returns from 30 various stocks and estimates mixtures of Gaussian distributions. Mixtures of Gaussian distribution is a conventional way of generalizing a given family of distributions, including heavy-tailed distribution which is the aim of the study.

Here the system innovation in the system equation is assumed to follow a two mixture of Gaussian distributions, denote by $N(0, \sigma_s^2) + I(t) \cdot N(0, \sigma_l^2)$, where the index function $I(t)$ follows a Bernoulli distribution, occasionally taking one with a chance of $P_l$, otherwise zero. Without loss of generality, $\sigma_l^2$ is larger than $\sigma_s^2$. Other model assumptions in the standard Kalman filter are maintained.

It is a well-known fact that Kalman filter with non-Gaussian distribution has no simple, tractable analytical solution as we have seen in the standard Kalman filter in equation (3.7). However, numerical solutions are available for various extended Kalman filter models. A conventional, commonly applied technique for numerical solutions in non-Gaussian Kalman filter, or Bayesian filtering problem in general, is sequential Monte Carlo (SMC) sampling (Kitagawa (1987), Doucet, Godsill, and Andrieu (2000), Doucet, De Freitas, and Gordon (2001)).

The implementation of SMC sampling in the case here follows,

1. First, the prior distribution of $x_{t-1}$ is generated by simulation. Here I denote each simulated sample with subscript "j", $\{x_{t-1}^j\}_{j=1}^{j=M}$, where the sample size $M$ is set to 40000 to ensure reliable numerical estimates.

2. Second, in order to generate the prior distribution of $x_t$, the system equation as well as another randomly drawn variables are employed; First, the sample of system innovations $\{u_j\}_{j=1}^{j=M}$ are randomly drawn using the mixture of Gaussian distribution, and then the simulated prior $\{x_{t-1}^j\}_{j=1}^{j=M}$ are added to the simulated system innovations to derive the prior distribution of $x_t$, denoted by $\{x_{p,t}^j\}_{j=1}^{j=M}$. Here the resulting

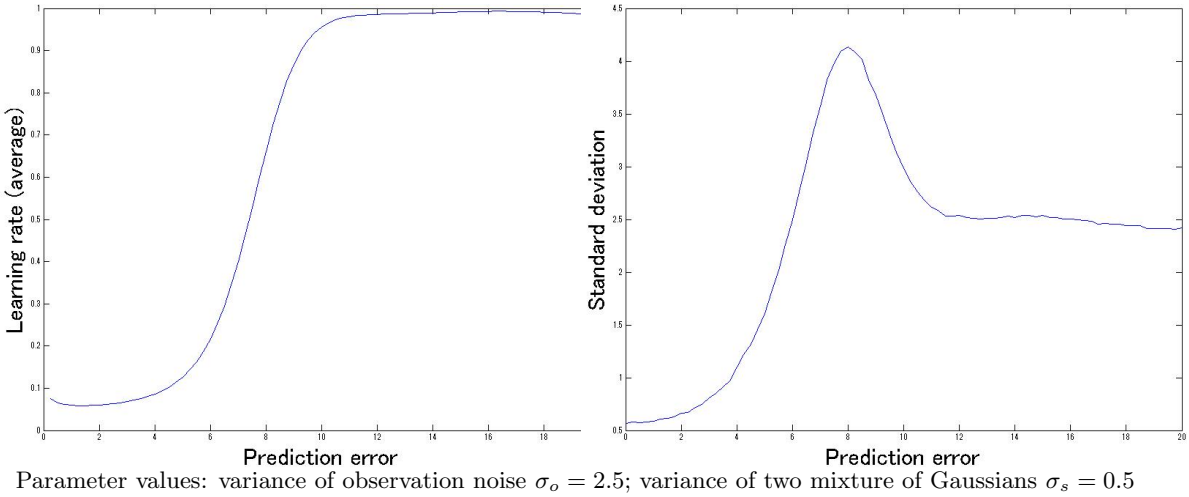
distribution is the prior distribution of $x_t$ before observing the new evidence $y_t$.

3. Finally, in order to derive the posterior distribution $f(x_t|y_t)$, an importance sampling is employed; resample the prior distribution of $x_t$ ($\{x^j_{p,t}\}^{j=M}_{j=1}$) with weight of the probability distribution $f(y_t|x^j_{p,t})$, denoting the resulting sample by $\{x^j_t\}^{j=M}_{j=1}$. Since the observation error $e_t$ follows a Gaussian distribution, the probability distribution $f(y_t|x^j_{p,t})$ has an analytically tractable expression and is computed with little difficulty.

Then, the resulting numerical distribution $\{x^j_t\}^{j=M}_{j=1}$ follows Bayesian posterior distribution $f(x_t|y_t)$, given the prior distribution $f(x_{t-1})$ and the evidence $y_t$. Moreover, recursively, the numerical distribution $\{x^j_t\}^{j=M}_{j=1}$ is employed as the prior distribution of the next stage, namely referred as "sequential" Monte Carlo sampling.

Figure 3.3: Learning under outlier risk in system innovation

Mean (left) and standard deviation (right) of posterior belief




Parameter values: variance of observation noise $\sigma_o = 2.5$; variance of two mixture of Gaussians $\sigma_s = 0.5$ and $\sigma_l = 23.5$; the probability of mixture of Gaussian with larger variance $P_l = 0.125$.

The left panel of Figure 3.3 is the plot of simulated learning rate $A_t$ (mean of the posterior belief) over the absolute value of prediction error under outlier risk in system innovations, computed using the SMC sampling. For simplicity of the simulation, the fundamental value

$x_t$ of the previous period is assumed to be known without uncertainty to eliminate the effect of the uncertainty in prior belief. The learning rate is a S-shape, upward function of the absolute value of the prediction error, implying that the less precise the prediction is, the more magnified the feedback of the prediction error is into the value update process. The shape is remarkably different from the one in the standard Gaussian world in which the functional form was a constant value with respect to prediction error. The intuition behind the S-shaped learning rate is quite reasonable; once a prediction error is large, people optimally attribute the most part of the error to the system innovation since larger changes are easier to occur in system innovation due to its heavier-tailed distribution than that of the observation noise.

The right panel of the Figure 3.3 is the plot of the standard deviation of the posterior belief on the state variable $x_t$ over prediction error, as a measure of the subjective uncertainty of the belief. Contrarily to the mean value of the belief, the standard deviation is not a monotonic function of prediction error, rather it is a hump-shaped; It monotonically increases to reach a plateau with prediction error being approximately 8, then decrease to be a constant value with prediction error being approximately 12 or more. This result suggests that the prediction of the fundamental value is the most inaccurate and unstable at a middle size of forecast error, rather than extremely large deviations, perhaps as people cannot attribute the principal source of the prediction error well, to either system innovation or observation noise.

### 3.3.3 Optimal learning under outlier risk in observation noise

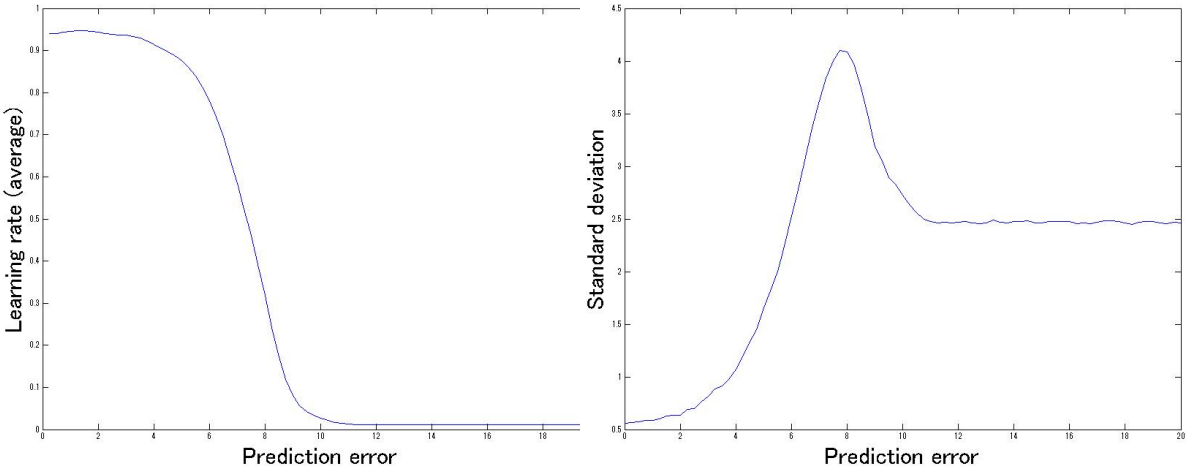
Contrary to traditional finance theories, behavioral finance has proposed a theory with empirical evidence that stock prices and their fundamental values occasionally diverge significantly, often referred as price anomalies (De Bondt and Thaler (1989), Shiller (2003)). A price deviation from the fundamental value is temporal and the price tends to revert to its fundamental value predictably in a short term or a long term. Although the earlier

studies on price reversions were mainly focused on relatively long-term reversions (Bondt and Thaler (1985)), Bremer and Sweeney (1991) also find the evidence of the short-term reversions of stock prices associated with outliers of their price changes, suggesting that these outliers include significant price anomalies.

To allow occasional outliers in the price anomalies instead of fundamental values, here I assume that the observation noise in the observation equation follows a two mixture of Gaussian distributions as same as in the case of outliers in fundamental values, denote dy $N(0, \sigma_s^2) + I(t) \cdot N(0, \sigma_l^2)$, where again the index function $I(t)$ follows Bernoulli distribution, occasionally taking one with a probability of $P_l$, otherwise zero.

Figure 3.4: Learning under outlier risk in observation noise

Mean (left) and standard deviation (right) of posterior belief




Parameter values: variance of system innovation $\sigma_i = 2.5$; variance of two mixture of Gaussians $\sigma_s = 0.5$ and $\sigma_l = 23.5$; the probability of mixture of Gaussian with larger variance $P_l = 0.125$.

The SMC sampling of the previous model is also applicable to this case with little modifications. The left panel of Figure 3.4 shows the plot of simulated learning rate with respect to the absolute value of prediction error under outlier risk in observation error using SMC sampling, given that the fundamental value $x_t$ of the previous period is fully revealed for the sake of simplicity of the simulation. Indeed, the shape of the function has a distinc-

tive feature; The function of the learning rate is an inverse-S-shape, downward function of the absolute value of prediction error; The less precise the prediction is, the less the feedback of the prediction error is. Again, the shape is remarkably different from the one in Gaussian world (a constant value with respect to the value of prediction error). The inverse-S-shape function also differs from that in the case for outliers in system innovations case, whose learning rate was the S-shape upward function. In the world of outlier risk in price anomalies, people optimally attribute the prediction error to the observation error rather than the system innovations when they are faced with a large prediction error, since now the observation error has a heavier-tail in its distribution rather than that of the system innovation.

The right panel of the Figure 3.4 is the plot of the standard deviation of the posterior belief. Interestingly, the shape of the plot is almost identical to the case of system innovation with outliers while the learning rate curve is remarkably different; It has a hump-shaped having a maximum value with prediction error being around 8, and decrease to be a constant value. Again, the agent is maximumly uncertain on the source of the prediction error at the middle value, rather than extremely large prediction error.

## 3.4 Discussion

In the present study I show the optimal learning for price dynamics under outlier risk. Two kinds of outlier risk in price processes are considered here; A price process in which outliers occur as its fundamental value has changed, and a price process in which outliers occur even with little fundamental change as deviations from its fundamental value, rather they are due to observation noise. The latter is often referred as price anomaly in behavioral finance.

The two optimal learning models are characterized as Bayesian reinforcement-learning models using non-Gaussian Kalman filter, and are solved numerically with sequential Monte

Carlo sampling. The numerical solutions revealed several key features of the two optimal learning model, which are mainly summarized in their learning rate and prediction error; The learning rate with outlier risk in fundamental value is a monotonically increasing function of absolute value of prediction error, while the learning rate with outlier risk in observation noise is a monotonically decreasing function. Interestingly, the uncertainly of the learning is seemingly identical among the two, having a hump-shaped function of absolute value of prediction error.

Several further research directions should be considered, however, I believe one of the most fruitful would be investigation on neural substrates of outlier risk as less is known about the neurobiological foundations of outlier processing in humans. Only a few suggestions have been made in the literature (Yu and Dayan (2005), Aston-Jones and Cohen (2005), Sara (2009)), with links to neuromodulators like norepinephrine, or specific brain structures such as hippocampus, but little concrete is known. Moreover, neuroscientists have tended to investigate only outliers that are associated with fundamental/regime changes of environments, rather than outliers that are less associated. As often discussed in finance literature, outliers play a crucial role in financial markets. The profound knowledge on human outlier processing would help to understand human behavior in financial markets, perhaps uncovering the human suboptimal behavior for their price predictions.

# Bibliography


Ackerberg, D. (2003): "Advertising, Learning, and Consumer Choice in Experience Good Markets: A Structural Examination," *International Economic Review*, 44, 1007–1040.

Aguirregabiria, V., and A. Magesan (2011): "Identification and Estimation of Dynamic Games when Players' Beliefs are not in Equilibrium," mimeo., University of Toronto.

Armel, K., A. Beaumel, and A. Rangel (2008): "Biasing simple choices by manipulating relative visual attention," *Judgment and Decision Making*, 3(5), 396–403.

Armel, K., and A. Rangel (2008): "The impact of computation time and experience on decision values," *American Economic Review*, 98(2), 163–168.

Aston-Jones, G., and J. Cohen (2005): "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance," *Annu. Rev. Neurosci.*, 28, 403–450.

Bajari, P., and A. Hortaçsu (2005): "Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data," *Journal of Political Economy*, 113, 703–741.

Banks, J., and R. Sundarum (1992): "Denumerable-Armed Bandits," *Econometrica*, 60, 1071–1096.

Behr, A., and U. Pötter (2009): "Alternatives to the normal model of stock returns: Gaussian mixture, generalised logF and generalised hyperbolic models," *Annals of Finance*, 5(1), 49–68.

Bondt, W., and R. Thaler (1985): "Does the stock market overreact?," *The Journal of Finance*, 40(3), 793–805.

Boorman, E., T. Behrens, M. Woolrich, and M. Rushworth (2009): "How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action," *Neuron*, 62(5), 733–743.

Bremer, M., and R. Sweeney (1991): "The Reversal of Large Stock-Price Decreases," *The Journal of Finance*, 46(2), 747–754.

Brocas, I., J. Carrillo, S. Wang, and C. Camerer (2009): "Measuring attention and strategic behavior in games with private information," mimeo., USC.

Bromberg-Martin, E., M. Matsumoto, and O. Hikosaka (2010): "Dopamine in motivational control: rewarding, aversive, and alerting," *Neuron*, 68(5), 815–834.

Brown, A., C. Camerer, and D. Lovallo (2010): "To review or not to review? limited strategic thinking at the movie box office," *American Economic Journal: Microeconomics*.

Camerer, C., and T. Ho (1999): "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67(4), 827–874.

Camerer, C., and E. Johnson (2004): "Thinking about attention in games: backward and forward induction," *The Psychology of Economic Decisions, Volume Two: Reasons and Choices ed. by I. Brocas and J. Carrillo.*

Camerer, C., E. Johnson, T. Rymon, and S. Sen (1993): "Cognition and framing in sequential bargaining for gains and losses," *Frontiers of game theory*, pp. 27–47.

Caplin, A., and M. Dean (2008): "Economic Insights from 'Neuroeconomic' Data," *The American Economic Review*, 98(2), 169–174.

Caplin, A., M. Dean, P. Glimcher, and R. Rutledge (2010): "Measuring Beliefs and Rewards: a Neuroeconomic Approach," *Quarterly Journal of Economics*, 125, 923–960.

Chan, T., and B. Hamilton (2006): "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 114, 997–1040.

Charness, G., and D. Levin (2005): "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect," *American Economic Review*, 95, 1300–1309.

Choi, J., D. Laibson, B. Madrian, and A. Metrick (2009): "Reinforcement learning and savings behavior," *The Journal of Finance*, 64(6), 2515–2534.

Cochrane, J. (2005): *Asset pricing*. Princeton University Press.

Costa-Gomes, M., and V. Crawford (2006): "Cognition and behavior in two-person guessing games: An experimental study," *American Economic Review*, 96(5), 1737–1768.

Costa-Gomes, M., V. Crawford, and B. Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69(5), 1193–1235.

Craik, K. (1943): *The nature of explanation*. Cambridge University Press.

Crawford, G., and M. Shum (2005): "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 73, 1137–1174.

Crawford, V., and N. Iriberri (2007): "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, 75(6), 1721–1770.

Dayan, P., and A. Yu (2003): "Uncertainty and learning," *IETE Journal of Research*, 49(2/3), 171–182.

De Bondt, W., and R. Thaler (1989): "Anomalies: A mean-reverting walk down Wall Street," *The Journal of Economic Perspectives*, 3(1), 189–202.

Doucet, A., N. De Freitas, and N. Gordon (2001): *Sequential Monte Carlo methods in practice*. Springer Verlag.

Doucet, A., S. Godsill, and C. Andrieu (2000): "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and computing*, 10(3), 197–208.

El-Gamal, M., and D. Grether (1995): "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of American Statistical Association*, 90, 1137–1145.

Erdem, T., and M. Keane (1996): "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15, 1–20.

Gabaix, X., D. Laibson, G. Moloche, and S. Weinberg (2006): "Costly information acquisition: Experimental analysis of a boundedly rational model," *The American Economic Review*, 96(4), 1043–1068.

Gentner, D., and A. Stevens (1983): *Mental models*. Lawrence Erlbaum.

Ghahramani, Z. (2001): "An Introduction to Hidden Markov Models and Bayesian Networks," *International Journal of Pattern Recognition and Artificial Intelligence*, 15, 9–42.

Gillen, B. (2011): "Identification and Estimation of Level-k Auctions," mimeo., Caltech.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17(3), 295–314.

Glimcher, P., C. Camerer, R. Poldrack, and E. Fehr (2008): *Neuroeconomics: decision making and the brain*. Academic Press.

Goldfarb, A., and M. Xiao (2011): "Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets," *American Economic Review*, 101(7), 3130–3161.

Grether, D. (1992): "Testing Bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, 17(1), 31–57.

Griffin, D., and A. Tversky (1992): "The weighing of evidence and the determinants of confidence," *Cognitive Psychology*, 24(3), 411–435.

Hampton, A., P. Bossaerts, and J. O'Doherty (2006): "The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans," *Journal of Neuroscience*, 26, 8360–8367.

Hampton, A., P. Bossaerts, and J. O'Doherty (2008): "Neural correlates of mentalizing-related computations during strategic interactions in humans," *Proceedings of the National Academy of Sciences*, 105(18), 6741.

Harrison, P., and C. Stevens (1971): "A Bayesian approach to short-term forecasting," *Operational Research Quarterly*, 22(4), 341–362.

Hikosaka, O., K. Nakamura, and H. Nakahara (2006): "Basal ganglia orient eyes to reward," *Journal of Neurophysiology*, 95(2), 567.

Ho, T., and X. Su (2010): "A Dynamic Level-k Model in Games," mimeo., University of California at Berkeley.

Holt, C. (1986): "Scoring-rule procedures for eliciting subjective probability and utility functions," *Bayesian inference and decision techniques, ed. by P. Goel and A. Zellner*, pp. 279–290.

Hossain, T., and R. Okui (2010): "The Binarized Scoring Rule of Belief Elicitation," Discussion paper.

Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, and C. Camerer (2005): "Neural systems responding to degrees of uncertainty in human decision-making," *Science*, 310(5754), 1680–1683.

Hsu, M., I. Krajbich, C. Zhao, and C. Camerer (2009): "Neural response to reward anticipation under risk is nonlinear in probabilities," *The Journal of Neuroscience*, 29(7), 2231–2237.

Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: a General Solution," *Journal of Econometrics*, 144, 27–61.

Johnson, E., C. Camerer, S. Sen, and T. Rymon (2002): "Detecting failures of backward induction: Monitoring information search in sequential bargaining," *Journal of Economic Theory*, 104(1), 16–47.

Jones, D., and J. Gittins (1972): *A dynamic allocation index for the sequential design of experiments*. University of Cambridge, Department of Engineering.

Kawagoe, R., Y. Takikawa, and O. Hikosaka (1998): "Expectation of reward modulates cognitive signals in the basal ganglia," *Nat Neurosci*, 1, 411–416.

Kitagawa, G. (1987): "Non-Gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, 82(400), 1032–1041.

Knoepfle, D., J. Wang, and C. Camerer (2009): "STUDYING LEARNING IN GAMES USING EYE-TRACKING," *Journal of the European Economic Association*, 7(2-3), 388–398.

Kon, S. (2012): "Models of stock returns–a comparison," *Journal of Finance*, 39(1), 147–165.

Krajbich, I., C. Armel, and A. Rangel (2010): "Visual fixations and the computation and comparison of value in simple choice," *Nature Neuroscience*, 13, 1292–1298.

Kruschke, J., and M. Johansen (1999): "A model of probabilistic category learning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1083.

Kuhnen, C., and B. Knutson (2008): "The Influence of Affect on Beliefs, Preferences and Financial Decisions," MPRA Paper 10410, University Library of Munich, Germany.

Lauwereyns, J., K. Watanabe, B. Coe, and O. Hikosaka (2002): "A neural correlate of response bias in monkey caudate nucleus," *Nature*, 418, 413–417.

Lohse, G. (1997): "Consumer eye movement patterns on yellow pages advertising," *Journal of Advertising*, pp. 61–73.

Malkiel, B. (1999): *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

——— (2003): "The efficient market hypothesis and its critics," *The Journal of Economic Perspectives*, 17(1), 59–82.

Meinhold, R., and N. Singpurwalla (1983): "Understanding the Kalman filter," *American Statistician*, 37(2), 123–127.

Nyarko, Y., and A. Schotter (2002): "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70, 971–1005.

Payne, J., J. Bettman, and E. Johnson (1993): *The adaptive decision maker*. Cambridge University Press.

Payzan-LeNestour, E., and P. Bossaerts (2011): "Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings," *PLoS Computational Biology*, 7(1), 1704–1711.

Preuschoff, K., P. Bossaerts, and S. Quartz (2006): "Neural differentiation of expected reward and risk in human subcortical structures," *Neuron*, 51(3), 381–390.

Preuschoff, K., B. Marius Hart, and W. Einhauser (2011): "Frontiers: Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making," *Frontiers in Decision Neuroscience*, 5.

Rangel, A. (2009): "The computation and comparison of value in goal-directed choice," *Neuroeconomics: Decision Making and the Brain, ed. by PW Glimcher, CF Camerer, E. Fehr, and R. Poldrack*, pp. 425–440.

Rescorla, R., and A. Wagner (1972): "Variations in the Effectiveness of Reinforcement and Nonreinforcement," *New York: Classical Conditioning II: Current Research and Theory, Appleton-Century-Crofts*.

Reutskaja, E., R. Nagel, C. Camerer, and A. Rangel (2011): "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study," *American Economic Review*, 101(2), 900–926.

Rushworth, M., and T. Behrens (2008): "Choice, uncertainty and value in prefrontal and cingulate cortex," *Nature Neuroscience*, 11(4), 389–397.

Sara, S. (2009): "The locus coeruleus and noradrenergic modulation of cognition," *Nature Reviews Neuroscience*, 10(3), 211–223.

Schultz, W., P. Dayan, and P. Montague (1997): "A neural substrate of prediction and reward," *Science*, 275(5306), 1593.

Shiller, R. (2003): "From efficient markets theory to behavioral finance," *The Journal of Economic Perspectives*, 17(1), 83–104.

Shimojo, S., C. Simion, E. Shimojo, and C. Scheier (2003): "Gaze bias both reflects and influences preference," *Nature Neuroscience*, 6(12), 1317–1322.

Sokol-Hessner, P., M. Hsu, N. Curley, M. Delgado, C. Camerer, and E. Phelps (2009): "Thinking like a trader selectively reduces individuals' loss aversion," *Proceedings of the National Academy of Sciences*, 106(13), 5035.

Strahilevitz, M., T. Odean, and B. Barber (2011): "Once burned, twice shy: How naïve learning, counterfactuals, and regret affect the repurchase of stocks previously sold," *Journal of Marketing Research*, 48(SPL), 102–120.

Sutton, R., and A. Barto (1998): *Reinforcement Learning*. MIT Press.

Symmonds, M., P. Bossaerts, and R. Dolan (2010): "A behavioral and neural evaluation of prospective decision-making under risk," *The Journal of Neuroscience*, 30(43), 14380–14389.

Taleb, N. (2008): *The black swan*. Random House, Inc.

Thaler, R., and C. Sunstein (2008): *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Wang, J., M. Spezio, and C. Camerer (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, 100(3), 984–1007.

Wells, C. (1996): *The Kalman filter in finance*, vol. 32. Springer.

Wunderlich, K., U. Beierholm, P. Bossaerts, and J. O'Doherty (2011): "The human prefrontal cortex mediates integration of potential causes behind observed outcomes," *Journal of Neurophysiology*, 106(3), 1558–1569.

Yu, A., and P. Dayan (2005): "Uncertainty, neuromodulation, and attention," *Neuron*, 46(4), 681–692.

Zhang, J., M. Wedel, and R. Pieters (2009): "Sales Effects of Attention to Feature Advertisements: a Bayesian Mediation Analysis," *Journal of Marketing Research*, 46, 669–681.

# Appendix A

# Appendix of Chapter 2

## A.1 Details of optimal Bayesian learning model

Here we provide more details about the simulation of the optimal learning and decision rules from Section 2.2.1. First we introduce some notation and describe the information structure and how Bayesian updating would proceed in the reversal learning context. Let $(Y_t, S_t, R_t)$ denote the actions, state, and rewards. Furthermore, let $Q$ denote the $2 \times 2$ Markov transition matrix for the state $S_t$, corresponding to Matrix (2.2).

Let $B_t^*$ denote the *prior belief* that $S_t = 1$, at the beginning of period $t$, while $\tilde{B}_t^*$ denotes the *posterior* belief that $S_t = 1$, at the end of period $t$, after taking action $Y_t$ and observing reward $R_t$. The relationship between $B_t^*$ and $\tilde{B}_t^*$ is given by Bayes rule:

$$\tilde{p}_t = P(S_t = 1|p_t, R_t, Y_t) = \frac{p_t \cdot f(R_t|S_t = 1, Y_t)}{p_t \cdot f(R_t|S_t = 1, Y_t) + (1 - p_t) \cdot f(R_t|S_t = 2, Y_t)}$$

Combining this with $Q$, we obtain the period-by-period transition for the prior beliefs $B_t^*$:

$$\begin{bmatrix} B_{t+1}^* \\ 1 - B_{t+1}^* \end{bmatrix} = Q \cdot \begin{bmatrix} \tilde{B}_t^* \\ 1 - \tilde{B}_t^* \end{bmatrix} = Q \cdot \begin{bmatrix} P(S_t = 1|B_t^*, R_t, Y_t) \\ 1 - P(S_t = 1|B_t^*, R_t, Y_t) \end{bmatrix} \tag{A.1}$$

Next we describe a dynamic Bayesian learning model for the reversal-learning environment.

As in the experiments, we consider a finite (25 period) horizon, with $t = 1, \ldots T = 25$. Each subject's objective is to choose sequence of actions to maximize expected rewards:

$$\max_{i_1, i_2, \ldots, i_T} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

The state variable in this model is $B_t^*$, the beliefs at the beginning of each period. Correspondingly, the Bellman equation is:

$$\begin{aligned} V_t(B_t^*) &= \max_{Y_t \in \{1,2\}} \left\{\mathbb{E}\left[R_t + V_{t+1}(B_{t+1}^*)|Y_t, B_t^*\right]\right\} \\ &= \max_{Y_t \in \{1,2\}} \left\{\mathbb{E}\left[R_t|Y_t, B_t^*\right] + \mathbb{E}_{R_t|Y_t,B_t^*}\mathbb{E}_{B_{t+1}^*|B_t^*,Y_t,R_t} V_{t+1}(B_{t+1}^*)\right\} \end{aligned} \tag{A.2}$$

Above, the expectation $E_{B_{t+1}^*|B_t^*,Y_t,R_t}$ is taken with respect to Eq. (A.1), the law of motion for the prior beliefs, while the expectation $E_{R_t|Y_t,B_t^*}$ is derived from the assumed distribution of $(R_t|Y_t, \omega_t)$ via

$$P(R_t|Y_t, B_t^*) = B_t^* \cdot P(R_t|Y_t, \omega_t = 1) + (1 - B_t^*) \cdot P(R_t|Y_t, \omega_t = 2).$$

While we have not been able to derive closed-form solutions to this dynamic optimization problem, we can compute the optimal decision rules by backward induction. Specifically, in the last period $T = 25$, the Bellman equation is:

$$V_T(B_T^*) = \max_{Y_t \in \{1,2\}} E\left[R_t|Y_t, B_T^*\right]. \tag{A.3}$$

We can discretize the values of $B_T^*$ into the finite discrete set $\mathcal{B}$. Then for each $B \in \mathcal{B}$, we can solve Eq. (A.3) to obtain the period-$T$ value and choice functions $\hat{V}_T(B)$ and $\hat{y}_T^*(B) = \text{argmax}_i \mathbb{E}[R_t|i, B]$ for each value of $B \in \mathcal{B}$. Subsequently, proceeding backwards, we can obtain the value and choice functions for periods $t = T-1, T-2, \ldots, 1$. These choice functions are plotted in Figure 2.2.

## A.2 Details on model fitting and belief estimation in different learning models

In Section 2.5, we compared belief dynamics in the nonparametric model ($X^*$) with counterparts in other two benchmark learning models, the Bayesian belief ($B^*$) and the valuation in the reinforcement-learning model ($V_b - V_g$). Here we provide additional details for how the beliefs for each of the three models were computed.

### A.2.1 Belief dynamics $X^*$ in the nonparametric model

The values of $X^*$, the belief process in our nonparametric learning model, were obtained by maximum likelihood. For each block, using the estimated choice and measurement probabilities, as well as the learning rules, we chose the path of beliefs $\{X_t^*\}_{t=1}^{25}$ which maximized $P(\{X_t^*\} \,|\, \{Z_t, R_t\})$, the conditional ("posterior") probability of the beliefs, given the observed sequences of eye-movements and rewards. Because

$$P(\{X_t^*, Z_t\} \,|\, \{Y_t, R_t\}) = P(\{X_t^*\} \,|\, \{Z_t, R_t\}) \cdot P(\{Z_t\} \,|\, \{Y_t, R_t\}),$$

where the second term on the RHS of the equation above does not depend on $X_t^*$, it is equivalent to maximize $P(\{X_t^*, Z_t\} \,|\, \{Y_t, R_t\})$ with respect to $\{X_t^*\}$. Because of the Markov structure, the joint log-likelihood factors as:

$$\log L(\{X_t^*, Z_t\}|\{Y_t, R_t\}) = \sum_{t=1}^{24} \log \left[P(Z_t|X_t^*)P(X_{t+1}^*|X_t^*, R_t, Y_t)\right] + \log(P(Z_{25}|X_{25}^*)). \tag{A.4}$$

We plug in our nonparametric estimates of $P(Z|X^*)$ and $P(X_{t+1}^*|X_t^*, R_t, Y_t)$ into the above likelihood, and optimize it over all paths of $\{X_t^*\}_{t=1}^{25}$ with the initial condition restriction $X_1^* = 2$ (beliefs indicate "not sure" at the beginning of each block). To facilitate this optimization problem, we derive the optimal sequence of beliefs using a dynamic-programming

(Viterbi) algorithm; cf. Ghahramani (2001).

In the above, we treated the choice sequence $\{Y_t\}$ as exogenous, and left the choice probabilities $P(Y_t|X_t^*)$ out of the log-likelihood function (A.4) above. By doing this, we essentially ignore the implied correlation between beliefs and choices in estimating beliefs. This was because, given our estimates that $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$ in Table 2.3, maximizing with respect to these choice probabilities would leads to estimates of beliefs $\{X^*\}$ which closely coincide with observed choices; we wished to avoid such an artificially good "fit" between the beliefs and observed choices.

For robustness, however, we also estimated the beliefs $\{X^*\}$ including the choice probabilities $P(Y_t|X_t^*)$ in the likelihood function. Not surprisingly, the correlation between choices and beliefs $\text{Corr}(Y_t, X_t^*) = 0.99$, and in practically all periods, the estimated beliefs and observed choices coincided (ie. $X_t^* = Y_t$). However, we felt that this did not accurately reflect subjects' beliefs.

### A.2.2 Bayesian learning model

The learning and decision rules for the Bayesian model were described and computed in Section 2.2.1, with additional details provided in Appendix A.1. The sequence of Bayesian beliefs $B_t^*$ is obtained from Eq. (A.1) and evaluated at the observed sequence of choices and rewards $(Y_t, R_t)$.

### A.2.3 Reinforcement-learning model

We employ a variant of the TD (Temporal-Difference)-Learning models (Sutton and Barto (1998) , Section 6) in which action values are updated via the Rescorla-Wagner rule (Rescorla and Wagner (1972)). The value updating rule for a one-step TD-Learning model is given by:

$$V_{Y_t}^{t+1} \longleftarrow V_{Y_t}^t + \alpha\delta_t. \tag{A.5}$$

where $Y_t$ denotes the choice taken in trial $t$, $\alpha$ denotes the learning rate, and $\delta_t$ denotes the "prediction error" $\delta_t$ for trial $t$, defined as:

$$\delta_t = R_t - V_{Y_t}^t, \tag{A.6}$$

the difference between $R_t$ (the observed reward in trial $t$) and $V_{Y_t}^t$ (the current valuation). In trial $t$, only the value for the chosen alternative $Y_t$ is updated; there is no updating of the valuation for the choice that was not taken.

$P_c^t$, the current probability of choosing action $c$, is assumed to take the conventional "soft-max" (i.e., logit) form with the temperature parameter $\tau$:

$$P_c^t = e^{V_c^t/\tau} / \left[\sum_{c'} e^{V_{c'}^t/\tau}\right] \tag{A.7}$$

We estimated the parameters $\tau$ and $\alpha$ using maximum likelihood. For greater model flexibility, we allowed the parameter $\alpha$ to differ following positive vs. negative rewards. The estimates (and standard errors) are:

$$\begin{aligned} &\tau = 0.2729 \ \ (0.0307) \\ &\alpha \text{ for positive reward } (R_t = 2) = 0.7549 \ \ (0.0758) \\ &\alpha \text{ for negative reward } (R_t = 1) = 0.3333 \ \ (0.0518). \end{aligned} \tag{A.8}$$

We plug in these values into Eqs. (A.5), (A.6), and (A.10) to derive a sequence of valuations $\left\{V_t^* \equiv V_b^t - V_g^t\right\}$. The choice function (Eq. (A.7)) can be rewritten as a function of the difference $V_t^*$; i.e., the choice probability for the blue slot machine is,

$$P_b^t = \frac{e^{(V_b^t - V_g^t)/\tau}}{1 + e^{(V_b^t - V_g^t)/\tau}} = \frac{e^{V_t^*}/\tau}{1 + e^{V_t^*/\tau}} \tag{A.9}$$

and $P_g^t = 1 - P_b^t$. Hence, $V_t^*$ plays a role in the TD-Learning model analogous to the belief measures $X_t^*$ and $B_t^*$ from, respectively, the nonparametric and Bayesian learning models.

### A.2.4 Pseudo-Bayesian learning model

A pseudo-Bayesian learner uses Bayes rule to update her belief (as in the Optimal Bayesian model), but her choices are determined (suboptimally) by the "softmax" rule, as in reinforcement learning:

$$P_c^t = e^{B^{*t}_c/\tau} / \left[ \sum_{c'} e^{B^{*t}_{c'}/\tau} \right] \tag{A.10}$$

The maximum-likelihood estimate of $\tau$ is 0.2176 with bootstrapped standard error of 0.0138.

### A.2.5 Win-stay model

The final model is a simple behavioral heuristic. If subjects choose a slot machine and receive the positive reward $R_t = 1$, they repeat the choice in the next period with probability $1 - \delta$ (and switch to the other choice with probability $\delta$). If they choose a slot machine but obtain the negative reward $R_t = -1$, they switch to the other slot machine in the next trial with probability $1 - \epsilon$.

We estimated the parameters $\delta$ and $\epsilon$ using maximum likelihood. The estimates we obtained from the data were:

$$\delta = 0.1268 \;\; (0.0142); \quad \epsilon = 0.4994 \;\; (0.0213). \tag{A.11}$$

## A.3 Details on discretization of eye movements

In this section, we present additional discussion on the discretization of the eye-movement measure, and some evidence that a three-valued discretization (which we used in our preferred empirical specifications) is sufficient to capture most of the variation in this measure.

We start by assessing the discretization of $Z_t$ using a statistical approach based on the condi-

tion number of the matrix containing the sample conditional probabilities of the discretized values of $(Z_t|Z_{t-1})$. The intuition is straightforward: if we discretize into an excessive number of points, the matrix containing the discretized conditional distribution of $(Z_t|Z_{t-1})$ will be singular, reflecting the redundancy of information in the overly discretized values of $Z_t$.

Furthermore, and more importantly, our identification assumptions imply that the rank of the matrix $G_{z_t|z_{t-1}}$ is related to the dimension of the unobserved beliefs $X^*$; thus, a proper discretization of $Z_t$ is required to obtain reasonable estimates of beliefs. Formally, given a discretization of $Z$ into $K$ categories, the largest possible rank is

$$rank\left(G_{z_t|z_{t-1}}\right) = \begin{cases} K & \text{if } K \leq \dim(X^*) \\ \dim(X^*) & \text{if } K > \dim(X^*) \end{cases}.$$

In other words,

$$G_{z_t|z_{t-1}} \text{ is } \begin{cases} \text{nonsingular} & \text{if } K \leq \dim(X^*) \\ \text{singular} & \text{if } K > \dim(X^*) \end{cases}.$$

Therefore, the dimension of $X^*$ is the largest discretization $K$ with which $G_{z_t|z_{t-1}}$ is still nonsingular.

The condition number of a matrix measures how "close to singular" a given matrix is, with a larger condition number indicating a less well-behaved matrix.[1] In this exercise, we discretize $Z_t$ into number of points ranging from 2 to 6, and computed the condition number for the matrix of sample conditional probabilities $G(Z_t|Z_{t-1})$ in each case. Table A.1 shows the condition numbers for each case, along with block-bootstrap estimates of the percentiles of their sampling distribution.

As the results show, the big jump in condition number occurs between 3 and 4; the sample

[1] Formally, the condition number of a matrix measures the sensitivity of the solution of a system of linear equations to errors in the data, which depends on the invertibility of the matrix containing the coefficients of the linear equations. Values of condition number near 1 indicate a well-conditioned matrix, while a larger condition numbers suggests that a matrix is close to singular.

Table A.1: Condition number of matrix $G_{z_t|z_{t-1}}$

| | Dimension | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ |
| original sample | 3.3586 | 10.6571 | 48.1296 | 292.3680 | 198.9212 |
| mean | 3.3627 | 10.7541 | 176.1842 | 981.4516 | 1302.2 |
| minimum | 3.2270 | 6.3758 | 14.3674 | 29.7191 | 30.9333 |
| $5th$ percentile | 3.2270 | 8.2144 | 22.2867 | 66.7 | 62.1 |
| $25th$ percentile | 3.2594 | 9.5753 | 33.9472 | 107.9 | 109.6 |
| median | 3.2980 | 10.7541 | 49.7292 | 175.1 | 185.4 |
| $75th$ percentile | 3.3898 | 12.1294 | 88.8227 | 401.6 | 407.5 |
| $95th$ percentile | 3.6331 | 15.1768 | 345.9043 | 1985.0 | 2530.2 |
| maximum | 3.6331 | 22.6323 | 39013 | 273180 | 222780 |

condition number jumps almost five-fold from 10.7 to 48.1, while the estimated sampling distribution also blows up, with the 95th percentile jumping from 15.2 to 345.9. This offers some statistical confirmation for the three-point discretization of the eye movement measure $Z_t$ used in our empirical analysis.

Besides this formal statistical evidence, we also present the raw histogram of the undiscretized eye-movement measure $\tilde{Z}_t$ in Figure A.1. It is apparently trimodal, with peaks at -1, 0 and 1, suggesting that a three-value discretization of $Z_p$ indeed captures most of its variation. In the empirical work, we use the following three-value discretization as follows:

$$Z_t = \begin{cases} 1 & \text{if } \tilde{Z}_t < -\sigma_z \\ 2 & \text{if } -\sigma_z \leq \tilde{Z}_t \leq \sigma_z \\ 3 & \text{if } \sigma_z < \tilde{Z}_t \end{cases} \tag{A.12}$$

where $\sigma_z$ denotes a discretizing constant. As the baseline, we set $\sigma_z = 0.20$. However, we do not find any difference in the estimation results either qualitatively nor significantly if we vary $\sigma_z$ from 0.05 to around 0.40, suggesting that the model is robust for different classifications. Table A.2 shows the sample frequencies of the discretized measure $Z_t$ for three different values of $\sigma_z$.

Figure A.1: Histogram of undiscretized eye-movement measure $\tilde{Z}_t$

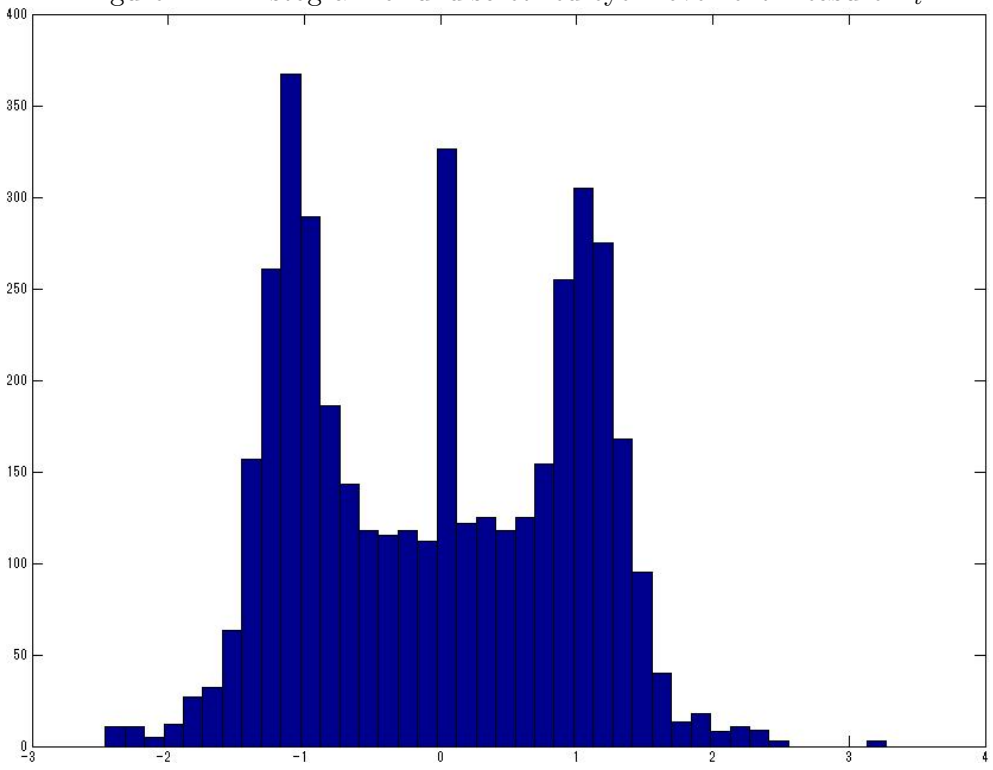


Table A.2: Correlations between $(Y, \tilde{Z})$ in different subsamples

| | Size | Corr$(Y, \tilde{Z})$ |
|---|---|---|
| Full sample | 4200 | 0.7647 |
| **$\sigma_z$ = 0.20 (baseline):** | | |
| $Z = 1$ (green) | 1887 | 0.2845 |
| 2 (not sure) | 540 | 0.2156 |
| 3 (blue) | 1773 | 0.1706 |
| **$\sigma_z$=0.05:** | | |
| $Z = 1$ (green) | 2015 | 0.3223 |
| 2 (not sure) | 255 | -0.0599 |
| 3 (blue) | 1930 | 0.2346 |
| **$\sigma_z$=0.40:** | | |
| $Z = 1$ (green) | 1725 | 0.1462 |
| 2 (not sure) | 869 | 0.2777 |
| 3 (blue) | 1606 | 0.0991 |

Note: $\tilde{Z}_t$ refers to the undiscretized eye-movement measure, as defined in Eq. (2.3), and $Z$ refers to the discretized version, as defined in Eq. (A.12).

Moreover, Table A.2 also shows the correlations between $Y$ and $\tilde{Z}$, broken up into the three ranges of $\tilde{Z}$ corresponding to the three discretized values $Z \in \{1, 2, 3\}$. Although the correlation between $(Y, \tilde{Z})$ in the whole sample is 0.7647, the correlations within each of the three ranges of $\tilde{Z}$ drop significantly, ranging from even negative values to values around 0.30. Because most of the variation in choices is *across* the different discretized values of $Z$, rather than within these values, it appears the three-valued discretization is sufficient.

## A.4 Conditional serial correlation in eye movements

In this section we assess more formally one critical part of Assumption 3, which is that the eye-movement measures are serially independent, conditional on beliefs. That is, $P(Z_t|Z_{t-1}, X_t^*) = P(Z_t|X_t^*)$. Since this exclusion restriction plays a crucial role in pinning down the values

of the beliefs, we assess it by estimating an alternative model in which we do not impose this assumption. In this alternative model, the "measurement probabilities" are given by the conditional distribution of $f(Z_t|X_t^*, Z_{t-1})$. In the remainder of this section, we describe how this expanded model is estimated.

Consider the joint density $f(Z_t, Y_t, Z_{t-1}, Z_{t-2})$, which is solely a function of variables observed in the data. Following the approach taken in Section 2.3.1 of the main text, we can factor this density as follows:

$$\begin{aligned}
& f(Z_t, Y_t|Z_{t-1}, Z_{t-2}) \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f(Z_t, Y_t, X_t^*, X_{t-1}^*|Z_{t-1}, Z_{t-2}) \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f(Z_t|Y_t, X_t^*, X_{t-1}^*, Z_{t-1}, Z_{t-2}) f(Y_t|X_t^*, X_{t-1}^*, Z_{t-1}, Z_{t-2}) f(X_t^*, X_{t-1}^*|Z_{t-1}, Z_{t-2}) \\
= & \sum_{X_t^*} f(Z_t|X_t^*, Z_{t-1}) f(Y_t|X_t^*) \sum_{X_{t-1}^*} f(X_t^*, X_{t-1}^*|Z_{t-1}, Z_{t-2})
\end{aligned}$$

For a fixed $z_{t-1}$, we have

$$f(Z_t, Y_t|z_{t-1}, Z_{t-2}) = \sum_{X_t^*} f(Z_t|X_t^*, z_{t-1}) f(Y_t|X_t^*) f(X_t^*|z_{t-1}, Z_{t-2}).$$

Technically, for any fixed $Y_t = y_t$ and $Z_{t-1} = z_{t-1}$, then, we can write the above in matrix notation as:

$$\mathbf{A}_{y_t, Z_t|z_{t-1}, Z_{t-2}} = \mathbf{B}_{Z_t|X_t^*, z_{t-1}} \mathbf{D}_{y_t|X_t^*} \mathbf{C}_{X_t^*|z_{t-1}, Z_{t-2}}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all $K \times K$ matrices, and $\mathbf{D}$ is a $K \times K$ diagonal matrix. These are defined

$$\begin{aligned}
\mathbf{A}_{y, Z_t|z_{t-1}, Z_{t-2}} &= \left[ f_{Y_t, Z_t|Z_{t-1}, Z_{t-2}}(y_t, i|z_{t-1}, j) \right]_{i,j} \\
\mathbf{B}_{Z_t|X_t^*, z_{t-1}} &= \left[ f_{Z_t|X_t^*, Z_{t-1}}(i|k, z_{t-1}) \right]_{i,k} \\
\mathbf{C}_{X_t^*|z_{t-1}, Z_{t-2}} &= \left[ f_{X_t^*|Z_{t-1}, Z_{t-2}}(k|z_{t-1}, j) \right]_{k,j}
\end{aligned}$$

The key eigendecomposition equation, analogous to Eq. (2.7) for the simpler model, becomes

$$\mathbf{A}_{y_t,Z_t|z_{t-1},Z_{t-2}}\mathbf{G}^{-1}_{Z_t|z_{t-1},Z_{t-2}} = \mathbf{B}_{Z_t|X_t^*,z_{t-1}}\mathbf{D}_{y_t|X_t^*}\mathbf{B}^{-1}_{Z_t|X_t^*,z_{t-1}}$$

where[2]

$$\mathbf{G}_{Z_t|z_{t-1},Z_{t-2}} = \left[f_{Z_t|Z_{t-1},Z_{t-2}}(i|z_{t-1},k)\right]_{i,k}$$

Therefore, we can apply this eigendecomposition to estimate $f(Z_t|X_t^*, z_{t-1})$ for each value of $z_{t-1}$. To assess whether we need to allow for conditional serial correlation in $Z_t$, we can compare whether the estimated probabilities $f(Z_t|X_t^*, z_{t-1})$ differ in $z_{t-1}$, i.e.,

$$f(Z_t|X_t^*, \widetilde{z}_{t-1}) \stackrel{?}{=} f(Z_t|X_t^*, \overline{z}_{t-1}).$$

The estimates of the probabilities $f(Z_t|X_t^*, z_{t-1})$ for $z_{t-1} = 1, 3$ are presented in Table A.3.[3] As the results show, the estimates of these probabilities are quite similar across different values of $z_{t-1}$. This suggests that conditional serial correlation in eye movements is not a major concern, and supports Assumption 3 underlying our empirical model.

## A.5 Belief updating and choices following "unsure" belief state

In Section 2.5.2 of the main text, we present some evidence, based on comparing the eye-movement measures to the beliefs from the Bayesian model, that eye movements were noisy measurements of beliefs, and not just of choices. Here, we consider another assessment of this crucial assumption which underlies our empirical model.

Here, we exploit that fact that, in our model, beliefs (and eye movements) take more values than choices. We consider what happens when beliefs are "unsure"; that is, when beliefs

[2]Note that the invertibility of $\mathbf{G}_{Z_t|z_{t-1},Z_{t-2}}$ is testable for each $z_{t-1}$.

[3]We were not able to estimate $f(Z_t|X_t^*, z_{t-1} = 2)$ because we observed too few observations with $z_{t-1} = 2$.

Table A.3: Measurement probabilities: $P(Z_t|X_t^*)$

| | $Z_{t-1} = 1$, $(N = 1748)$ | | |
|---|---|---|---|
| | $X_t^* = 1$ | $X_t^* = 2$ | $X_t^* = 3$ |
| $Z_t = 1$ | 0.8522 | 0.2138 | 0.0797 |
| | (0.1060) | (0.1508) | (0.0472) |
| $Z_t = 2$ | 0.0923 | 0.4523 | 0.1257 |
| | (0.0649) | (0.1264) | (0.0508) |
| $Z_t = 3$ | 0.0555 | 0.3340 | 0.7945 |
| | (0.0546) | (0.1285) | (0.0679) |
| | $Z_{t-1} = 2$, $(N = 487)$ | | |
| | Insufficient sample size. | | |
| | $Z_{t-1} = 3$, $(N = 1629)$ | | |
| | $X_t^* = 1$ | $X_t^* = 2$ | $X_t^* = 3$ |
| $Z_t = 1$ | 0.7844 | 0.1706 | 0.0574 |
| | (0.0950) | (0.1170) | (0.0513) |
| $Z_t = 2$ | 0.0732 | 0.5398 | 0.1744 |
| | (0.0553) | (0.2023) | (0.1160) |
| $Z_t = 3$ | 0.1425 | 0.2879 | 0.7682 |
| | (0.0697) | (0.2019) | (0.1378) |

note 1: Cutoff for the three-value discretization is 0.2,

note 2: Standard errors (in parentheses) computed across 1500 bootstrap resamples.

$X_t^*$ take the intermediate value of 2. Since eye movements play a crucial role in pinning down beliefs, if eye movements are just a noisy measure of choices, then choices should be similar following the "unsure" state ($X_t^* = 2$) than following the "sure" states ($X_t^* = 1$ or $X_t^* = 3$). However, if eye movements contain extra information beyond that contained in the choices, then we should find that belief updating and choice behavior following the "unsure" state is different than that following the "sure" states. The goal is to show that the "unsure" state matters for both belief updating and decision-making.[4]

First we show that beliefs update different following the unsure state than following a sure state. To do this, we perform a joint test that the probabilities in the leftmost column (corresponding to beliefs following belief-congruent choices) of each transition matrix in Table 4 differs from the middle column. We construct the test statistic as follows. Let $\vec{L}$ (resp. $\vec{M}$) denote the left-hand (resp. middle) column of a matrix in Table 4, omitting the bottom element. The test statistic is the quadratic form $(\vec{L} - \vec{M})'\Sigma^{-1}(\vec{L} - \vec{M})$, where $\Sigma$ is the variance-covariance matrix of $(\vec{L} - \vec{M})$ which was computed by bootstrap (as was all the estimates in Table 4).

Asymptotically, under the null hypothesis of no differences between the columns, this statistic is distributed according to a $\chi^2$-distribution, with two degrees of freedom. The corresponding $p$-values are given in Table A.4. The $p$-values are all small; the first two $p$-values imply that the "not sure" and "green" belief states are distinct, while the last two indicate that the "not sure" and "blue" states differ. Hence, the unsure state ($X_t^* = 2$) matters, in the sense that beliefs in the following period $X_{t+1}^*$ are statistically different when $X_t^*$ is a sure ("blue", "green") state versus an unsure state.

However, beliefs are unobservable. How does the unsure state affect future *observed* choices? To do this, we used our estimation results to compute the conditional distributions

$$Y_{t+1}|X_t^*, Y_t, R_t = \sum_{i=1}^{3} (Y_{t+1}|X_{t+1}^* = i) \cdot (X_{t+1}^* = i|X_t^*, Y_t, R_t).$$

[4]We are grateful to a referee for this suggestion.

Table A.4: Tests of belief updating following unsure state
$H_0$: $P(X^*_{t+1}|X^*_t = 1, Y_t, R_t) = P(X^*_{t+1}|X^*_t = 2, Y_t, R_t)$

| $(R_t, Y_t)$: | $(1,1)$ | $(2,1)$ | $(1,2)$ | $(2,2)$ |
|---|---|---|---|---|
| $p$-value: LH=middle[a] | 0.016 | 0.121 | 0.032 | 0.072 |

[a]Each entry contains the $p$-value under the null hypothesis that the leftmost and middle columns in the corresponding matrix in Table 5 have the same values. Under the null hypothesis, the test statistic has an asymptotic $\chi^2$ distribution with two degrees of freedom.

The conditional distribution $Y_{t+1}|X^*_t, Y_t, R_t$ describes how choices in period $t$ are made, conditional of beliefs, choices, and rewards in period $t$. This distribution is given in Table A.5. As before, we want to test whether the "unsure" state ($X^*_t = 2$) has distinctive effects on observed choices. To do this, we test, as before, whether the leftmost and middle columns of each matrix in Table A.5 are the same. The $p$-values under the null that these two columns are the same are also reported in Table A.5. These $p$-values are small, indicating scant evidence favoring the null hypothesis; while we cannot reject the null hypothesis at conventional significance levels in two of the four cases (corresponding to $(Y_t = 1, R_t = 1)$ and $(Y_t = 2, R_t = 2)$), the small $p$-values do favor the hypothesis that the leftmost and middle columns are different. Thus, we also find that the "unsure" state is important in predicting choices, which implies that the eye movements $Z_t$ contain more information than is contained in choices alone. This lends support to our modeling assumption that eye movements are noisy measures of beliefs.

Table A.5: How current beliefs affect future choices
The conditional probabilities $Y_{t+1}|X_t^*, Y_t, R_t$ computed from estimation results

$P(Y_{t+1}|X_t^*, y, r)$, $r$ =1(lose), $y$ =1(green)

| $X_t^*$ | 1(green) | 2 (not sure) | 3(blue) |
|---|---|---|---|
| $Y_{t+1}$ = 1 (green) | 0.5675 | 0.4445 | 0.3551 |
| 2 (blue) | 0.4325 | 0.5555 | 0.6449 |

$P(Y_{t+1}|X_t^*, y, r)$, $r$ =2(win), $y$ =1(green)

| $X_t^*$ | 1(green) | 2 (not sure) | 3(blue) |
|---|---|---|---|
| $Y_{t+1}$ = 1 (green) | 0.8777 | 0.7731 | 0.8909 |
| 2 (blue) | 0.1223 | 0.2270 | 0.1091 |

$P(Y_{t+1}|X_t^*, y, r)$, $r$ =1(lose), $y$ =2(blue)

| $X_t^*$ | 3(blue) | 2 (not sure) | 1(green) |
|---|---|---|---|
| $Y_{t+1}$ = 2 (blue) | 0.5653 | 0.3527 | 0.2806 |
| 1 (green) | 0.4347 | 0.6473 | 0.7195 |

$P(Y_{t+1}|X_t^*, y, r)$, $r$ =2(win), $y$ =2(blue)

| $X_t^*$ | 3(blue) | 2 (not sure) | 1(green) |
|---|---|---|---|
| $Y_{t+1}$ = 2 (blue) | 0.8795 | 0.8114 | 0.8270 |
| 1 (green) | 0.1205 | 0.1980 | 0.1731 |

$H_0$: $P(Y_{t+1}|X_t^* = 1, Y_t, R_t) = P(Y_{t+1}|X_t^* = 2, Y_t, R_t)$

| $(Y_t, R_t)$: | $(1,1)$ | $(1,2)$ | $(2,1)$ | $(2,2)$ |
|---|---|---|---|---|
| $p$-value: LH=middle[a] | 0.109 | 0.091 | 0.022 | 0.163 |

[a]Each entry contains the $p$-value under the null hypothesis that the leftmost and middle columns in the corresponding matrix in the above table have the same values. Under the null hypothesis, the test statistic has an asymptotic standard normal distribution.

## A.6 Additional figures

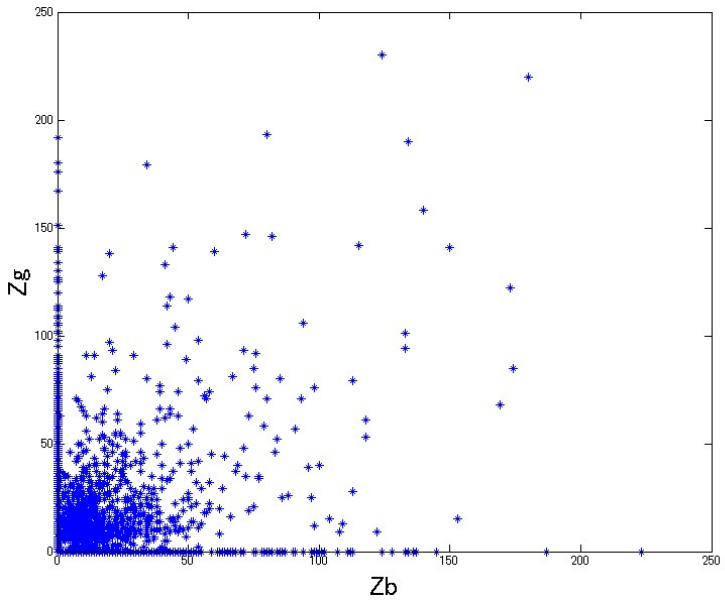



Figure A.2: Scatter plot of $Z_b$ (fixation on blue) and $Z_g$(fixation on green)
Both $Z_b$ and $Z_g$ are reported in $2 \times 10^{-2}$ seconds.

## A.7 Experimental instruction

### Screen 1

This experiment lasts for approximately 60 minutes, consisting of 8 blocks. Each block consists of 25 trials.

*Press SPACEBAR to continue.

### Screen 2

In each trial, you are asked to choose one of 2 slot machines to play a gamble. One is presented on the left side of the screen, and the other is presented on the right side. One is blue and the other is green. The location of each slot machine will vary from trial to trial.

*Press SPACEBAR to continue

### Screen 3

If you choose the left slot machine, press "LEFTARROW".

If you choose the right slot machine, press "RIGHTARROW".

Please ALWAYS put your fingers on "LEFTARROW" key and "RIGHTARROW" key.

*Press SPACEBAR to continue.

### Screen 4

IMPORTANT;

It is important that you pay attention to the visual stimuli presented on the screen. In particular, please look at the fixation cross when it appears. The software will not allow

you to move on until the eyetracker has confirmed that you have focused on the fixation cross for an appropriate amount of time. Also remember to keep your eyes on the pictures of the slot machines while they are on the screen. You are free to look back and forth between them, but just be sure that you are always looking at one or the other.

To insure this, please do not look at the keyboard at any time. Instead, keep your eyes on the screen. To help you do this, you should keep your fingers on "LEFTARROW" and "RIGHTARROW" keys at all times.

*Press SPACEBAR to continue.

**Screen 5**

In each trial, ONE and ONLY ONE of the 2 slot machines is GOOD.

If the chosen slot machine is GOOD, then the probability of winning $0.5 is 70% and the probability of losing $0.5 is 30%.

If the chosen slot machine is BAD, then the probability of winning $0.5 is 40% and the probability of losing $0.5 is 60%.

*Press SPACEBAR to continue.

**Screen 6**

At the beginning of each block, you do not know which slot machine is good/bad. At that time, both the blue slot machine and the green slot machine are good with an equal probability.

*Press SPACEBAR to continue.

**Screen 7**

But, when each trial ends, a good slot machine might switch to be bad (and vice versa) with a small probability (15%) although good one tends to remain good with a large probability (85%). For example, a blue slot machine that was good in a trial might switch to be bad in the next trial with 15% chance.

*Press SPACEBAR to continue.

**Screen 8**

However, please note that when a slot machine becomes bad, the other slot machine ALWAYS becomes good. ALWAYS ONE and ONLY ONE of the 2 slot machines is good, and the other is bad. The position of each machine on screen will vary randomly; the COLOR is the only indicator of whether the slot machine is good/bad.

*Press SPACEBAR to continue.

**Screen 9**

Right after you choose a slot machine in each trial, the reward for the CHOSEN slot will be presented. (Reward for the other slot will NOT be revealed.)

*Press SPACEBAR to continue.

**Screen 10**

Your best strategy is simple: Choose the slot machine that you think is good in the trial, taking into account the possibility of good-bad slot machine switch between the blue/green.

The money you earn in this experiment will be paid to you with show-up fee after the experiment. (If negative, it will be subtracted from show-up fee.)

*Press SPACEBAR to continue.

### Screen 11

Practice

In order to get used to the software, you will now play 15 practice trials. (Note: the money you earn in the practice session will NOT be paid for you). Please note that in the actual experiment, one block consists of 25 trials.

*Press SPACEBAR to continue.

### Screen 12

[A BLOCK STARTS.]

A block is going to begin...

Please ALWAYS put your fingers on "LEFTARROW" and "RIGHTARROW" and wait. Please note that, at the beginning of the new block, the blue/green slot can be good with an equal probability.

*Press SPACEBAR to continue.

### Screen 13

[A TRIAL STARTS.]

After a fixation(+), you will see 2 slot machines. If you choose the left slot machine, press "LEFTARROW". If you choose the right slot machine, press "RIGHTARROW".

*Please keep looking at the fixation(+) when presented. The screen will change soon.

**Screen 14**

[TWO SLOT MACHINES ARE PRESENTED TO BE CHOSEN.]

**Screen 15**

Wait for the next trial... Please put your fingers on "LEFTARROW" and "RIGHTARROW" and wait. Please note that good-bad slot machine switch might occur between the blue/green with 15% chance.

[END OF A TRIAL. REPEAT THE TRIAL 25 TIMES IN A BLOCK(15 FOR PRACTICE BLOCK).]

**Screen 16**

This concludes the block.

So far, you finished [1,2,3,4,5,6,7,8] / 8 blocks.

Feel free to rest for a minute if you need a break.

*Press SPACEBAR when you are ready to continue on with the experiment.

[END OF A BLOCK.]

**Screen 17**

[SHOWN ONLY AT THE END OF PRACTICE BLOCK.]

This is the end of practice. Do you have any questions? (Please ask the experimenter if you do now.) If you want to go through the instruction and practice again, press "1".

If you are ready to start the experiment, press "0".

[THEN REPEAT 8 BLOCKS(SCREEN 12 - 16).]

### Screen 18

This concludes this experiment. Thank you for participating! Your final earning is [TOTAL PAYMENT]($).

# Appendix B

# Appendix of Chapter 3

## B.1 The experimental design

In this section, I describe the details of the experiment whose data is presented in Section 3.2. The experiment is performed by Bossaerts' research group as a pilot experiment for their own fMRI study[1]. The experiment were run over several weeks in June 2012. They used 15 subjects, recruited from Caltech Social Science Experimental Laboratory (SSEL) subject pool consisting of undergraduate/graduate students, post-doctoral students, each performing for 800 trials of "Target prediction task" described below. Subjects were paid a fixed show-up fee ($14), in addition to the amount won during the experiment, which was $18.7 on average.

### B.1.1 Target position prediction task

In the task, subjects will see a Target (red dot) that moves along a circle (Fig. 3.1)[2], in both clockwise and counterclockwise direction. On the same circle, subjects will also

[1]The author is engaged in the project as a lab member of Bossaerts7 research group. Actually the experiment is designed in the author's collaborative work with Peter Bossaerts.

[2]Since a fMRI study is the primary purpose of the study of Bossaerts group, the experimental design is optimized for fMRI imaging, rather than exactly depicting financial markets, such as circle structure and slider input design. However, the data generating processes behind the task are exactly same as the model premises in the previous sections.

observe a Robot (blue dot). The task will be to control the Robot by changing how much it responds to Target movements. Underneath the circle is a slider (Fig. 3.1). Moving the cursor towards "1" increases the adjustment of the Robot towards the Target's last move. At "1", the Robot catches up fully with the Target's position. Moving the cursor towards "0" decreases the Robot's adjustment. At "0", the Robot stays put, unresponsive to the Target's last move.

Subjects' task is to ensure that the Robot is always as close as possible to the Target's subsequent position. The task proceeds as follow;

1. At the beginning of a trial, subject will observe the Target move to a new place.
2. After the Target moved, the circle segment between the Target and Robot is highlighted with yellow (Fig. 3.1). Then, subjects have a chance to change the Robot's adjustment for tracking the Target using a mouse. When the mouse moves, the slider underneath the circle on the screen moves accordingly to let subjects see how much they are going to adjust.
3. Then, subjects see the Robot move toward the Target according to the adjustment they decided.
4. Then, we proceed to the subsequent trial. (One trial finishes in approximately 3 seconds.)

The whole task will consist of 2 blocks. Each block is subdivided in 2 consecutive sessions. Each session takes approximately 12 minutes, and is subdivided in 200 trials. Between each session, approximately 1.5 minutes resting period is scheduled.

The way the Target moves will differ across the two blocks. But within a block, the Target uses a fixed strategy to move along the circle. This strategy will never depend on what subjects do. They have two experimental treatments for the Target movements, namely fundamental and the anomaly treatments, and the two treatments are exactly same as the

model premises discussed in the previous section; The two processes are described with simple hidden state Markov processes, where the observation noise is Gaussian and state transition non-Gaussian (heavy-tailed, mixture of Gaussians) in the fundamental treatment, while observation noise is non-Gaussian and state transition Gaussian in the anomaly treatment.

More precisely, let $y_t$ and $x_t$ be the position of the Target and that of hidden state variable, respectively. As described in Section 2, $Y_t$ and $X_t$ follow the observational equation and system equation.

$$\begin{aligned} y_t &= x_t + e_t \quad \text{(observational equation)} \\ x_t &= x_{t-1} + u_t \quad \text{(system equation)} \end{aligned} \tag{B.1}$$

where $e_t$ is the observation noise, and $u_t$ is the system innovation.

In the Fundamental treatment, the observation noise $e_t$ follows a Gaussian distribution with mean zero and variance $\sigma^2$, denoted as $N(0, \sigma^2)$ while the system innovation follows a two mixture of Gaussian distributions, denote by $N(0, \sigma_s^2) + I(t) \cdot N(0, \sigma_l^2)$, where the index function $I(t)$ follows a Bernoulli distribution, occasionally taking one with a chance of $P_l$, otherwise zero. And vice versa in the Anomaly treatment.

### B.1.2 Incentive compatible payment design

The payment design of the task is designed to incentivize participants to express their true estimates for the subsequent position of the Target (a mean value of the posterior belief distribution). Roughly speaking, more precise their prediction is, more the probability of winning reward. In each session, three randomly chosen trials have a chance to be rewarded. And in each of these trials, subjects have a chance to win two dollars. The probability of winning is determined by the accuracy of their navigating Robot R to the subsequent

position of Target $T_{(t+1)}$, as follows,

$$\text{Probability (win)} = \max[0, 100 - 0.04 \cdot (T_{(t+1)} - R)^2](\%) \tag{B.2}$$

Where the unit for T and R is degree.

#### B.1.2.1 Proof for the incentive compatibility of the payment design

Here I show the outline of the proof of the incentive compatibility of the payment design in which subjects honestly announce the mean value of their posterior belief distribution.

Incentive compatible mechanism design for belief elicitation with non-risk neutral agents has not been developed yet (Holt (1986)). Recently Hossain and Okui (2010) propose an incentive compatible mechanism design for eliciting various statistics on the belief distribution, using a stochastic payment design. The bottom line of the design is that, instead of manipulating the amount of payment according to the error of predictions, manipulating the probability of winning (and pay a fixed amount, say \$1) enables experimenters to ignore concave shape of utility functions and treat as if the agent is risk neutral. The following proof follows Hossain and Okui (2010)).

Let $\hat{X}$ be an announced value for $X$ by an agent, and let $F(X)$ be the agent's belief distribution on $X$.

The payment design here is,

- Pay \$1 with probability of $g(X, \hat{X})$, where $g(X, \hat{X}) = 1 - a \cdot (X - \hat{X})^2$. In the equation $a$ could take any number [3].
- Pay \$0 with probability of $1 - g(X, \hat{X})$

[3] Any value of $a$ satisfies incentive compatibility, however, the loss function $a \cdot (X - \hat{X})^2$ has to take positive numbers in the support of X, in which $g(X, \hat{X})$ takes positive numbers.

Then, the expected utility of the task is,

$$\begin{aligned} & E_X[E_{g(X,\hat{X})}U] \\ & = E_X[g(X,\hat{X}) \cdot u(1) + (1 - g(X,\hat{X})) \cdot u(0)]. \end{aligned} \tag{B.3}$$

Normalization $u(0) = 0$ derives,

$$\begin{aligned} & = E_X[g(X,\hat{X}) \cdot u(1)] \\ & = u(1) \cdot E_X[g(X,\hat{X})] \\ & = u(1) \cdot E_X[1 - a(X - \hat{X})^2] \\ & = u(1) - a \cdot u(1) \cdot E_X[(X - \hat{X})^2]. \end{aligned} \tag{B.4}$$

The utility is maximized when $\hat{X}$ is a mean of the distribution of $\tilde{X}$. The rest of the proof is as same as the case for Proper Scoring Rules for a risk neutral agent. ∎

### B.1.3 Experimental instruction

The experiment will last approximately 60 minutes. It will consist of 2 blocks. Each block is subdivided in 2 consecutive sessions. Each session takes approximately 12 minutes, and is subdivided in 200 trials.

#### Task overview

In this game, you will see a Target that moves along a circle (Fig. 2.7), in both clockwise and counterclockwise direction. On the same circle, you will also observe a Robot. Your task will be to control the Robot by changing how much it responds to Target movements. Underneath the circle is a slider (Fig. 2.7 and Fig. B.1). Move the cursor towards "1" to increase the adjustment of the Robot towards the Target's last move. At "1", the Robot

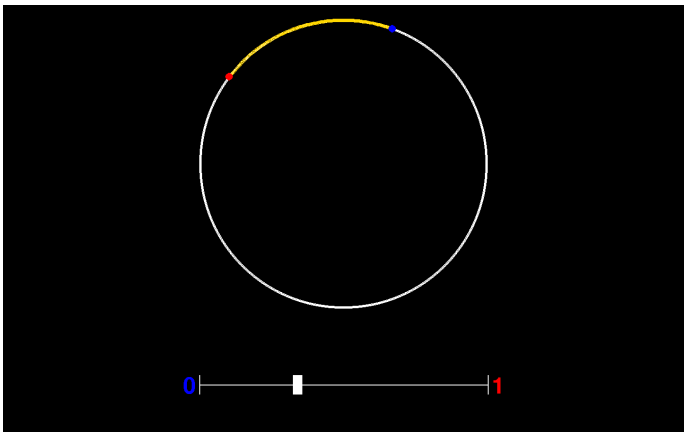



Figure B.1: Target prediction task (highlighted)

catches up fully with the Target's position. Move the cursor towards "0" to decrease Robot's adjustment. At "0", the Robot stays put, unresponsive to the Target's last move.

Your task is to ensure that the Robot is always as close as possible to the Target's SUBSEQUENT position. Occasionally, the Target might partly reverse its last movement.

The game is divided into "blocks". The way the Target moves (e.g., partly reverse its last movement) will generally differ across blocks. But within a block (two consecutive sessions), the Target uses a fixed strategy to move along the circle. This strategy will NEVER depend on what you do. Specifically, Target movements are autonomous, unaffected by Robot movements. For instance, the Target will not deliberately try to stay away from the Robot.

A block consists of two consecutive sessions. You will have a short break between sessions. When you finish the first block (session 1 and 2), you will move on to the second block (session 3 and 4). Again, The way the Target moves (e.g., partly reverse its last movement) will generally differ across blocks. But within a block (two consecutive sessions), the Target uses a fixed strategy to move along the circle.

At the end of the experiment, you will be able to win a show-up reward and bonus depending on your performance in the game. The chances of winning increase with your success in

making the Robot track the subsequent movements of the Target. The better you manage to track the Target in the subsequent trial, the higher the chances.

You will be given the opportunity to practice. After the experiment is over, we will ask a few questions about this task and your psychological and educational background. (The practice block consists of 100 trials and lasts for approximately 6 minutes.)

**Task protocol and interface**

Task Protocol:

1. (Fig. 3.1) At the beginning of a trial, you will see the Target $T_{(t-1)}$ moves to a new place ($T_{(t)}$) (Fig. 3.1). The red dot in the screen indicates the Target. The blue dot in the screen indicates the Robot.
2. After the Target moves, the circle segment between the Target and Robot is highlighted with yellow (Fig. B.1). Then, you have a chance to change the Robot's adjustment for tracking the Target, if you wish. (A slider is shown below on the screen. For details, see "input" below.).
3. You then see the Robot moves toward the Target, according to the adjustment you decided.
4. We then proceed to the subsequent trial soon. (Please note that, the above process goes quite rapidly; one trial finishes approximately in 3 seconds.)

In every trial, if you wish to change the Robot's adjustment for tracking the Target, you can change it by moving the cursor on the slider, right after the Target moves and the circle segment between the Target and the Robot is highlighted with yellow (Fig. B.1). Higher the value of the speed is, closer the Robot moves to the Target.

- At "1", the Robot fully catches up the Target.

- At "0", then the Robot stays put, unresponsive to the Target's last move.

You change the place of cursor with a track-ball (or a mouse in pilot/practice experiments). If you roll the tracking-ball to the left/right side, the cursor moves to the left/right side.

- The left-right direction of the slider bar may vary every 10 trials. In a certain trial, "1" might be located on the left side of the slider and "0" is on the right side (Fig. 3.1 and Fig. B.1), however, they might be reversed in another trial.
- When the slider is reversed, the slider becomes grey for a while, right before the next Robot movement.
- The starting value of the cursor is the value you decided in the last trial. (0.5 in the first trial.)

(Time limit) Please move the cursor on the slider within 2.25 seconds.

- The slider will be locked after 2.25 seconds (then the Robot moves).
- You may leave the cursor to the starting value if you do not wish to change the speed.

**Reward scheme**

- The following reward scheme may look a little bit complicated; all you need to know however, is that it incentivizes you to navigate the Robot close to what you think is the best estimates for the position of Target in the subsequent trial, $T_{(t+1)}$.
- As such, you really do not need to remember the details of the math of the scheme.

In each session, you will be rewarded for THREE randomly chosen trials.

In each of these trials, you have a chance to win TWO dollars. The probability of winning is determined by the accuracy of your navigating Robot R to the subsequent position of Target $T_{(t+1)}$, as follows,

$$\text{Probability (win)} = \max[0, 100 - 0.04 \cdot (T_{(t+1)} - R)^2](\%) \tag{B.5}$$

Where the unit for T and R is degree.

The following table (Table B.1) shows how the probability of winning changes with your Robot navigation. The closer your Robot is to the Target in the subsequent trial, the larger the probability of winning is.

Table B.1: Probability of winning the bonus of TWO dollars.

| Difference with Target and Robot | Chance of winning TWO dollars |
|---|---|
| 0 : | 100% |
| 20 degrees: | 84% |
| 30 degrees: | 64% |
| 40 degrees: | 36% |
| 50 degrees: | 0% |
| 60 degrees: | 0% |

* 30 degrees is equivalent to 1/12 of a circle, 60 degrees is 1/6.

To determine whether you won in a given trial, we generate a random variable with uniform distribution between 0 and 100. If the drawn value is smaller than the probability determined by the above equation, you will be awarded. Otherwise you earn nothing (for that trial).

Again, with this rule, your best strategy is simple: simply navigate the Robot to what you think is the best estimate of the subsequent position of the Target.

At the end of the whole experiment, we will pay the cumulative earnings from the selected trials in all blocks (plus the standard show-up reward).

Rewards are not paid for the practice block.

Thank you and good luck!