ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Hu, Yingyao; Kayaba, Yutaka; Shum, Matt

Working Paper

Nonparametric learning rules from bandit experiments: The eyes have it!

cemmap working paper, No. CWP15/10

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Hu, Yingyao; Kayaba, Yutaka; Shum, Matt (2010) : Nonparametric learning rules from bandit experiments: The eyes have it!, cemmap working paper, No. CWP15/10, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2010.1510

This Version is available at:
https://hdl.handle.net/10419/64791

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

cemmap
centre for microdata methods and practice

Nonparametric learning rules from bandit experiments: the eyes have it!

Yingyao Hu
Yutaka Kayaba
Matt Shum

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP15/10

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonparametric Learning Rules from Bandit Experiments: The Eyes have it!*

Yingyao Hu
Johns Hopkins

Yutaka Kayaba
Caltech

Matt Shum
Caltech

First draft: April 2010

Abstract

We estimate nonparametric learning rules using data from dynamic two-armed bandit (probabilistic reversal learning) experiments, supplemented with auxiliary eye-movement measures of subjects' beliefs. We apply recent econometric developments in the estimation of dynamic models. The direct estimation of learning rules differs from the usual *modus operandi* of the experimental literature. The estimated choice probabilities and learning rules from our nonparametric models have some distinctive features; notably that subjects tend to update in a non-smooth manner following positive "exploitative" choices (those made in accordance with current beliefs). Simulation results show how the estimated nonparametric learning rules fit aspects of subjects' observed choice sequences better than alternative parameterized learning rules from Bayesian and reinforcement learning models.

1 Introduction

How do individuals learn from past experience in dynamic choice environments? We address this question by presenting nonparametric estimates of subjects' learning rules in a dynamic two-armed bandit (probabilistic reversal learning) problem where subjects must repeatedly guess which of the two arms yields a (stochastically) higher reward. Auxiliary measures of subjects' eye movements as they make their choices are employed to "pin down" subjects'

*We are indebted to Antonio Rangel for his encouragement and for the funding and use of facilities in his lab.

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. Estimating the learning rules nonparametrically allows us to compare competing learning models in a manner quite distinctive than that taken in the existing literature.

Related literature There has been a large recent literature in industrial organization and marketing in which learning-based models of dynamic choice have been estimated structurally. Some representative papers include Ackerberg (2003), Erdem and Keane (1996), 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, depending on the laboratory setting, there is also the opportunity to observe "auxiliary" measures of subjects' beliefs, such as brain activity (cf. Yoshida and Ishii (2006) in the recent fMRI neuroscience literature) or eye movements (as in the present paper). Because of this data richness, researchers are able to consider more flexible learning rules, and to test the fully-rational Bayesian learning benchmark versus boundedly-rational, backward-looking "reinforcement learning" (RL) rules (cf. Sutton and Barto (1998)). This question has been tackled in the behavioral/experimental learning literature, including Charness and Levin (2005), Kuhnen and Knutson (2008), and Payzan and Bossaerts (2009). Particularly, RL 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 areas of the brain (cf. Schultz, Dayan, and Montague (1997)) for evidence from primates). Recently, RL models have also been used to explain some observed anomalies in savings and investment behavior (eg. Choi, Laibson, Madrian, and Metrick (2009), Odean, Strahilevitz, and Barber (2004)).[1]

[1] In the computational IO literature, such learning algorithms have also been used to ease the computational burden associated with dynamic equilibrium models, cf. Pakes and McGuire (2001), Imai, Jain, and Ching (2009).

The prevalent *modus operandi* in the behavioral/experimental literature has been to use the observed choice data from the experiment to calibrate parameters for competing learning models. Subsequently, the competing learning models are simulated, and verification is based upon comparing the simulated learning rules with the observed auxiliary belief measurements. For instance, Hampton, Bossaerts, and O'Doherty (2006) test between a Bayesian and reinforcement-learning model on the basis of two-armed bandit experiments supplemented with brain activity information from fMRI brainscans. Other papers utilizing a similar methodological framework include Behrens, Woolrich, Walton, and Rushworth (2007), Daw, O'Doherty, Dayan, Seymour, and Dolan (2006), Yoshida and Ishii (2006).

In this paper, we take a different approach. Instead of calibrating prespecified learning models, we use the observed experimental and auxiliary data to estimate, nonparametrically, subjects' learning rules, without imposing *a priori* functional forms on the learning rule. 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, including the Bayesian and reinforcement-learning models.

Moreover, we estimate not only the learning rules nonparametrically, but also the choice probabilities. Choice probabilities are key parameters in machine learning and decision neuroscience models (cf. Sutton and Barto (1998), Daw, O'Doherty, Dayan, Seymour, and Dolan (2006), Doya (2002)). Recently, researchers have worked on disentangling "exploitative" vs. "explorative" behavior, where the former refers to taking choices which yield high immediate rewards, while the latter refers to taking less familiar choices in order to gain information which might be more useful in the future.[2] Although parameterized models for exploration-exploitation behavior have been examined in several studies (cf. Daw, O'Doherty, Dayan, Seymour, and Dolan (2006)), to our knowledge, this research would be the first to examine choice behavior without imposing *a priori* functional forms on the choice probabilities.

Methodologically, this paper represents a novel application of econometric tools recently developed for the estimation of nonclassical measurement error models and dynamic discrete-choice models (Hu (2008), Hu and Shum (2008)). Because subjects' underlying beliefs are unobserved and also serially correlated over time, the learning model is a particular case of a nonlinear "hidden Markov" model, which are challenging to estimate (cf. Ghahra-

[2] Such a distinction is also present in the dynamic Bayesian learning framework, where explorative behavior is called "experimentation" (cf. Crawford and Shum (2005)).

mani (2001)). Our approach is to fit the learning model into a dynamic misclassification framework, in which the eye-movement measures play the role of "noisy measurements" of the underlying belief process. Clearly, this approach could also be used with fMRI data, which are richer in content than eye-tracking data. (Relatedly, Samejima, Doya, Ueda, and Kimura (2004) consider Bayesian estimation of a reinforcement learning model using sequential Monte Carlo ("particle filtering") methods.)

In Section 2, 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. We also describe our estimation procedure there. 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 reinforcement-learning models. Section 6 concludes.

2 Two-armed bandit learning (probabilistic reversal learning) experiment

The setup of the learning experiments is largely standard, and follows Hampton, Bossaerts, and O'Doherty (2006). We consider an experiment where subjects chooses between two actions, called "blue" and "green", where the rewards of these two actions are changing over time.

In each period t, a subject choose one of two auctions (which we call interchangeably "arms" or "slot machines" in what follows): $Y_t \in \{B, G\}$. Which of these arms is the "correct" one varies period-by-period, as described by the state variable $S_t \in \{1, 2\}$. The state variable 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 period t depends on the action taken, as well as (stochastically) on the current state: the good (bad) arm yields rewards

$$R_t = \begin{cases} 2 & \text{with prob } 0.7\ (0.4) \\ 1 & \text{with prob } 0.3\ (0.6) \end{cases}$$

The state evolves according to an exogenous binary Markov process, with transition prob-

abilities

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

Because S_t is not observed by subjects, and is serially-correlated over time, there is the opportunity for subjects to learn and update their beliefs about the current state on the basis of past rewards. The goal of the exercise in this paper is to infer subjects' learning (that is, belief updating) rule, on the basis of their observed choices.

2.1 Data

The experiments were run over several weeks time in November-December 2009. We used 21 subjects, recruited from the Caltech Social Science Experimental Laboratory (SSEL) subject pool consisting of undergraduate/graduate students, postdocs and community members, each playing for 200 rounds (8 blocks of 25 trials). For each subject, and each round t, we observe the data (Y_t, S_t, R_t). In Figure 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 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 for the "blue" slot machine on the computer screen.[3]

3 Econometric model

In this section, we describe our econometric model of dynamic decision-making in the two-armed bandit (probabilistic reversal learning) experiment described above, and also discuss the identification and estimation of this model. 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. 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

[3] 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 "ride side bias" which may be present (see discussion further below).

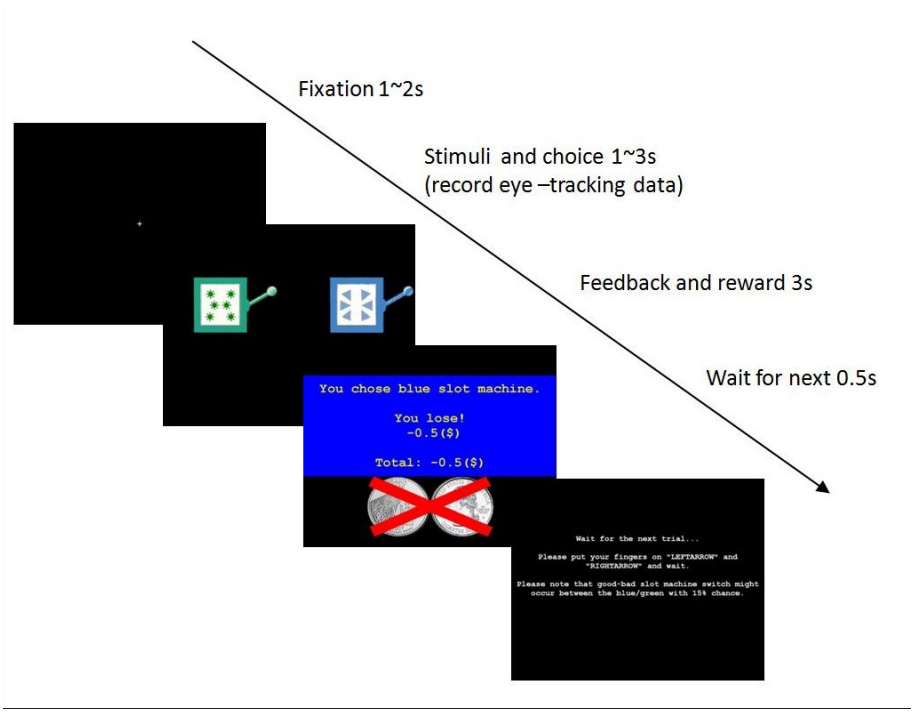

Figure 1: Timeline of a trial

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. After subjects make a choice (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.

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 $P(y|X^*)$ varies across different values of X_t^* (ie. beliefs affect actions).*

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 $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$.*

These two assumptions pose very little loss of generality, and hole for both the standard forward-looking Bayesian learning model (ie. as in Crawford and Shum (2005)) as well as most varieties of the backward-looking reinforcement-learning model.

Assumption 3 *The auxiliary measure Z_t is a noisy measure of beliefs X_t^*, with the measurement probabilities $P(Z_t|X_t^*)$. We assume that:*
(i) For all t, the $K \times K$ matrix $\mathbf{G}_{Z_t|Z_{t-1}}$, with entries $G(i,j) = P(Z_t = i|Z_{t-1} = j)$, is invertible.
(ii) $E[Z_t|X_t^*]$ is increasing in X_t^*.

The invertibility assumption 3(i) is made on the observed matrix $E_{Z_t|Z_{t-1}}$ with elements equal to the conditional distribution of $Z_t|Z_{t-1}$. 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.

The model can be easily extended to allow for conditional serial correlation in the auxiliary measure Z_t, ie. allowing for a law of motion $P(Z_t|X_t^*, Z_{t-1})$. For Z_t as a measure of eye-movements, as in this paper, the conditional independence assumption across trials appears reasonable, especially given the imposed fixation at the beginning and end of each trial (cf. Figure 1). However, for auxiliary measures in other settings (such as brain activity for fMRI studies), conditional dependence seems more realistic.

The final assumption justifies pooling the data across all subjects and trials for estimating the model:

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, trials, and rounds t.*

Given these assumptions, we next describe the nonparametric identification argument.

3.1 Nonparametric identification

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

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

Abusing terminology somewhat, we call this a "first-order Markov" property. This is because:

$$\begin{aligned} & f(Y_t, Z_t, X_t^*|Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \\ =& f(Y_t|X_t^*) \cdot f(Z_t|X_t^*) \cdot f(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1}) \\ =& f(Y_t, Z_t, X_t^*|Y_{t-1}, R_{t-1}, X_{t-1}^*). \end{aligned} \tag{2}$$

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

Consider the joint density $f(Z_t, Y_t|Z_{t-1})$, which is observed in the data. The main functions we want to identify are:
(i) $f(Y_t|X_t^*)$, the conditional choice probability;
(ii) the learning rule $f(X_t^*|X_{t-1}^*, Y_{t-1}, R_{t-1})$; and
(iii) $f(Z_t|X_t^*)$, the mapping between the auxiliary measure Z_t and the unobserved state 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 pause to note the difficulty of estimating this model. 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 $\mathbf{P(Y_t|X_t^*)}$ and measurement probabilities $\mathbf{P(Z_t|X_t^*)}$. We begin with the following factorization:

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

where the last equality applies 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.

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

$$\mathbf{A}_{y,Z_t|Z_{t-1}}\mathbf{G}^{-1}_{Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*}\mathbf{D}_{y|X_t^*}\mathbf{B}^{-1}_{Z_t|X_t^*}. \tag{3}$$

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 unobserved data sequence $\{Y_t, Z_t\}$. 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. (3) 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 is the eigendecomposition is unique, the representation in Eq. (3) 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 $\mathbf{G}$.

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

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

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

$$\mathbf{I}_{Z_{t+1},y,r,Z_t} = \mathbf{B}_{Z_{t+1}|X^*_{t+1}} \mathbf{J}_{X^*_{t+1},X^*_t,y,r} \mathbf{B}^T_{Z_t|X^*_t}$$

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

$$\mathbf{J}_{X^*_{t+1},X^*_t,y,r} = \mathbf{B}^{-1}_{Z_{t+1}|X^*_{t+1}} \mathbf{I}_{Z_{t+1},y,r,Z_t} \mathbf{B}^{T,-1}_{Z_t|X^*_t}. \tag{4}$$

This result implies that we can use two periods of data $Z_t, Y_t, R_t, Z_{t-1}, Y_{t-1}, R_{t-1}$ for the two steps.

3.2 Estimation

For the estimation, we assume that the variables Z_t and X_t^* are discrete, and take either two or three values. Since the eye-movement measure Z_t is continuous, we must discretize it for estimation. We postpone discussion of our exact discretization procedure until the next section, and full details are in the Appendix.

Our estimation procedure mimicks 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 $\mathbf{I}$ (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 $\mathbf{J}$ using the matrix manipulations in Eqs. (3) and (4).

One technical feature is that, because all the elements in the matrices of interest $\mathbf{J}$, $\mathbf{B}$, and $\mathbf{D}$ correspond to probabilities, they must take values $\in [0, 1]$. 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 the minimized the sum-of-squares corresponding to Eqs. (3) and (4), and explicitly restricted each element of the matrices to lie $\in [0, 1]$. However, as the estimates below show, this was not a frequent recourse; only a handful of the estimates reported below needed to be restricted in this manner.[4]

Before presenting the results, we present some Monte Carlo simulation results in Table 1, for simulated datasets around the same size as the datasets drawn from our experiments. These show 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.

Remark on eye-tracking measure Before presenting the estimation results, we pause to discuss the eye-tracking measure Z, and present some evidence showing that it is a plausible noisy measure of subjects' beliefs (and satisfies the monotonicity condition of Assumption 3 above).

Let Z_{pt} denote the undiscretized eye-movement measure, and Z_t the discretized measure. As discussed in the eye-tracking literature (cf. Krajbich, Armel, and Rangel (2007), Armel and Rangel (2008), Rangel (2008)), value computations and fixation durations in choice tasks are suggested to be closely related. Several seminal papers utilizing eye-tracking machines have confirmed that a longer fixation duration at an alternative implies a larger probability

[4] In addition, while the identification argument above was "cross-sectional" in nature, being based upon observations of three observations of $\{Y_t, Z_t, R_t\}$ per individual, in the estimation we exploited the long time series data we have for each subject, and pooled every "three time-continuous observations" $\{Y_{i,r,\tau}, Z_{i,r,\tau}, R_{i,r,\tau}\}_{\tau=t-1}^{\tau=t+1}$ across all subjects i, all rounds r, and all trials $\tau = 2, \ldots, 24$. 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 round. Under these assumptions, the ergodic theorem ensures that the (across time and subjects) sample frequencies used to construct the matrices $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{I}$ converge towards population counterparts.

Table 1: (FYI) Monte Carlo Results. (5000 iterations, median, ""= true value)

$P(Y_t \mid X_t^*)$		
X_t^*	1(green)	2(blue)
$Y_t = 1$	0.9501	0.0499
(green)	"0.9500"	"0.0500"
	(0.0165)	
2	0.0499	0.9501
(blue)	"0.0500"	"0.9500"

$P(Z_t \mid X_t^*)$		
X_t^*	1(green)	2(blue)
$Z_t = 1$	0.9003	0.0997
(green)	"0.9000"	"0.1000"
	(0.0150)	
2	0.0997	0.9003
(blue)	"0.1000"	"0.9000"

$P(X_{t+1}^* \mid X_t^*, y, r)$, r = 1(lose), y = 1(green)

X_t^*	1(green)	2(blue)
$X_{t+1}^* = 1$	0.3999	0.1746
(green)	"0.4000"	"0.1500"
	(0.0174)	(0.1508)
2	0.6001	0.8254
(blue)	"0.6000"	"0.8500"

$P(X_{t+1}^* \mid X_t^*, y, r)$, r = 2(win), y = 1(green)

X_t^*	1(green)	2(blue)
$X_{t+1}^* = 1$	0.7999	0.7113
(green)	"0.8000"	"0.7000"
	(0.0129)	(0.1287)
2	0.2001	0.2887
(blue)	"0.2000"	"0.3000"

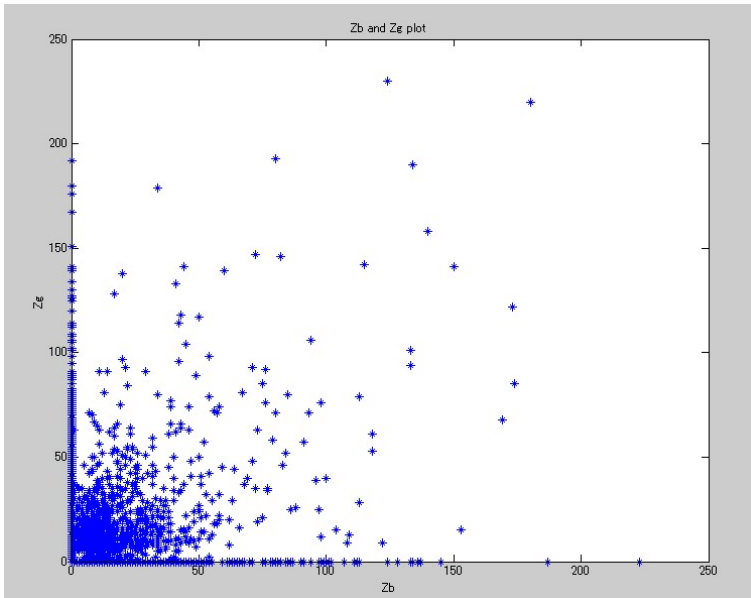

Figure 2: Scatter plot of Z_b (fixation on blue) and Z_g(fixation on green)

of choosing it. Following this literature, Z_{pt} is defined as,

$$Z_{pt} = \frac{(Z_{bt} - Z_{gt})}{RT_t}; \tag{5}$$

that is, $Z_{b(g)t}$ is the fixation duration at the blue (green) slot machine, and RT_t is the reaction time (ie. the time between the onset of the trial after fixation, and the subject's choice). Also in order to control for subject-specific heterogeneity, we normalize Z_{pt} across subjects by dividing by the subject-specific standard deviation of Z_{pt} (across all rounds for each subject), which we denote by "sid" in what follows.

Thus, Z_{pt} measures how much longer a subject look at the blue slot machine than the green one during the t-th trial, with a larger (smaller) value of Z_{pt} implying longer fixation time at the blue (green) slot machine. Figure 2 contains the scatter plot of Z_{bt} versus Z_{gt}, and Figure 3 is the histogram of Z_{pt}. The symmetric distribution around the 45-degree line in Figure 2, along with the symmetric shape around zero in Figure 3, indicates that there is no bias toward a certain color. Also we examine the existence of "right side bias". In the existing literature, it is often reported that human subjects exhibit a "right side bias", tending to gaze towards the right side more frequently. However, our experimental data contains no significant evidence of such a bias.

Table 2: Summary statistics for Y, R, Z_p, RT, Z

	green	blue
Y	2108	2092

	win	lose
R	2398	1802

	mean	median	upper 5%	lower 5%
Z_p	-0.0309	0	1.3987	-1.4091
RT	88.22	59.3	212.2	36.8

Sample size	21 subjects	168 blocks	4200 trials
Corr.(Y,Z_p)			0.7647

Z (after discretization with two values)

1(green, $Z_t < 0$)	2(blue, $Z_t \geq 0$)
2032	2168

Z (after discretization with three values)

sid.	1(green, $Z_t <$ -sid.)	2(not sure)	3(blue, $Z_t >$ sid.)
0.05	2015	255	1930
(baseline) 0.20	1887	540	1773
0.40	1725	869	1606

Moreover, this measure of Z_{pt} is well correlated with actual slot machine choices. Table 2 shows the summary statistics of Z_p. The correlation between Y_t (which =2(1) if blue(green) is chosen) and Z_{pt} is 0.7647, suggesting that Z_{pt} would be a good indirect measure of subjects' beliefs regarding whether the the blue slot machine is currently in the "good" state.

3.3 Estimation results

Two-value estimates In Table 3, we present estimates in the specification where X_t^* and Z_t are assumed to be binary variables taking values $\in \{1, 2\}$. The standard errors, shown in parentheses, were computed using bootstrap resampling (1000 iterations, resampled from all 168 blocks).

Starting from the top of the table, we see that the choice probabilities are reasonable, and very much aligned with beliefs. When $X_t^* = 1$ (associated with beliefs that "green is currently the good state"), then the green slot machine is pulled 98% of the time. Similarly, when $X_t^* = 2$, then the blue slot machine is chosen 94% of the time. Choice rules not completely aligned with beliefs are called "ϵ-greedy" rules in the learning literature; in backward-looking learning models (such as reinforcement learning, cf. Sutton and Barto (1998, pg. 28)), a deviation probability $\epsilon > 0$ is necessary to avoid getting "stuck" at suboptimal choices. Hence, the results here are consistent with an ϵ equal to around 5%.

The second panel in Table 3 contains the measurement probabilities. The estimates imply that beliefs closely track with the eye-movement measures, with (for instance) beliefs favoring green leading to longer gazes at the green slot machine on the computer screen around 92% of the time.

Finally, the remaining panels present the learning rule probabilities for all four configurations of $(R_t, Y_t) \in \{(1, 1), (2, 1), (1, 2), (2, 2)\}$. Note that the columns and rows are ordered differently across the panels, for ease of interpreting the results. Generally, the left column of each panel makes sense. Comparing the third and fourth panels in Table 3, we see that given the choice of "green" ($Y_t = 1$) and given beliefs in favor of green ($X_t^* = 1$), a higher reward leads to more intense updating of beliefs towards green in the next period; that is:

$$\begin{aligned} 0.87 &= P(X_{t+1}^* = 1 | X_t^* = 1, R_t = 2, Y_t = 1) \\ >> &P(X_{t+1}^* = 1 | X_t^* = 1, R_t = 1, Y_t = 1) = 0.54. \end{aligned}$$

Similarly, comparing the bottom two panels, we see that if the subject is predisposed towards

Table 3: Two-value estimates: Specification where X_t^* and Z_t are binary

$P(Y_t|X_t^*)$

X_t^*	1(green)	2(blue)
$Y_t = 1$	0.9756	0.0573
(green)	(0.0107)	(0.0167)
2	0.0244	0.9427
(blue)		

$P(Z_t|X_t^*)$

X_t^*	1(green)	2(blue)
$Z_t = 1$	0.9093	0.0888
(green)	(0.0156)	(0.0113)
2	0.0907	0.9112
(blue)		

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

X_t^*	1(green)	2(blue)
$X_{t+1}^* = 1$	0.5401	0.2950
(green)	(0.0284)	(0.1656)
2	0.4599	0.7050
(blue)		

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

X_t^*	1(green)	2(blue)
$X_{t+1}^* = 1$	0.8695	0.2471
(green)	(0.0192)	(0.2849)
2	0.1305	0.7529
(blue)		

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

X_t^*	2(blue)	1(green)
$X_{t+1}^* = 2$	0.5407	0.6836
(blue)	(0.0270)	(0.2621)
1	0.4593	0.3164
(green)		

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

X_t^*	2(blue)	1(green)
$X_{t+1}^* = 2$	0.9003	0.6146
(blue)	(0.0163)	(0.2484)
1	0.0997	0.3854
(green)		

blue ($X_t^* = 2$) then choosing blue $Y_t = 2$ and obtaining the higher reward $R_t = 2$ leads subjects to place a belief of 90% on "blue" the following period, vs. only 54% if this led to the lower reward $R_t = 1$.

On the other hand, the right columns in these panels are a bit puzzling. They indicate a great deal of state dependence in beliefs, when one chooses actions which are contrary to beliefs. For example, the third and fourth panels indicate that when $X_t^* = 2$ (so beliefs favor "blue"), but the subject chooses $Y_t = 1$ ("green"), then the updated beliefs are not affected much by the reward: with a high reward, beliefs switch to "green" ($X_{t+1}^* = 1$) with only 25% probability, but with a low reward, beliefs switched to "green" with the *slightly higher* probability of 30%, which is puzzling. Similarly, in the bottom two panels, when current beliefs favor "green" ($X_t^* = 1$), but the blue slot machine was chosen ($Y_t = 2$), then the probability that beliefs switched to "blue" ($X_{t+1}^* = 2$) is slightly higher following a low rather than high reward.

At face value, this suggests that subjects do not update their beliefs properly following "exploratory" (ie. contrary to belief) actions. However, as we will see now, these puzzling results are no longer so apparent when we allow beliefs to take three distinct values.

Three-value estimates Tables 4 and 5 present results from a specification where X_t^* is assumed to take three values $\{1, 2, 3\}$, and likewise Z_t is discretized to take these three values. 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".

Table 4 contains the estimates of the choice and measurement probabilities. 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 (ie. are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the "exploration probability" ϵ is smaller than in the two-value results, 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 towards "blue", with $Y_t = 2$ roughly 56% of the time. At the same time, the bottom panel indicates that when subjects are not sure, they tend to gaze in the middle of the screen, around 63% of the time.

The learning rule estimates are presented in Table 5. The results are similar to the two-value results, but some of the problems from those results disappear when we allow beliefs to take three values. The left columns show how beliefs are updated when "exploitative" choices

Table 4: Three-value estimates: Specification where X_t^* and Z_t take three values

Choice probabilities:

$P(Y_t|X_t^*)$

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

$P(Z_t|X_t^*)$

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

(ie. choices made in accordance with beliefs) are taken. We see that when current beliefs indicate "green" ($X_1^* = 1$) and green is chosen ($Y_t = 1$), beliefs are quite responsive to the reward: if $R_t = 1$ (the low reward), then beliefs stay at green with probability 57%, but if $R_t = 2$ (high reward), then this probability is much higher, at 89%. On the other hand, even after positive (ie. high reward) exploitative choices, beliefs may still update towards "blue" ($X_{t+1}^* = 3$) with an 11% chance, rather than sticking at the intermediate level $X_{t+1}^* = 2$. This non-smooth "extremal" updating is a distinctive feature of our learning rule estimates, and is consistent with optimal belief-updating in a probabilistic reversal context: even if the subject were completely sure that "green" after a high reward, she still must consider the possibility that the good state could change to "blue" by the next trial, due to stochastic state process. Moreover, as we will see below, this non-smooth updating following positive exploitative choices seems to be important in explaining choices which, through the lens of other standard learning models, appear "explorative" and contrarian.

The results in the right-most columns (describing belief updating follow "explorative" (contrarian) choices) are on the whole more sensible than in the two-value estimates. 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

Table 5: Three-value estimates: Specification where X_t^* and Z_t take three values

Learning Rule updating probabilities:

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

X_t^*	1(green)	2 (not sure)	3(blue)
$X_{t+1}^* = 1$	0.5724	0.3075	0.1779
(green)	(0.0694)	(0.0881)	(0.2257)
2	0.0000	0.3138	0.4002
(not sure)	(0.0662)	(0.1042)	(0.2284)
3	0.4276	0.3787	0.4219
(blue)	(0.0624)	(0.0945)	(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)
$X_{t+1}^* = 1$	0.8889	0.6621	0.8242
(green)	(0.0894)	(0.1309)	(0.2734)
2	0.0000	0.2702	0.1758
(not sure)	(0.0911)	(0.1297)	(0.1981)
3	0.1111	0.0678	0.0000
(blue)	(0.0340)	(0.0485)	(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)
$X_{t+1}^* = 3$	0.5376	0.2297	0.2123
(blue)	(0.0890)	(0.0731)	(0.1436)
2	0.0458	0.2096	0.1086
(not sure)	(0.0732)	(0.0958)	(0.1524)
1	0.4166	0.5607	0.6792
(green)	(0.0874)	(0.0968)	(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)
$X_{t+1}^* = 3$	0.8845	0.6163	0.6319
(blue)	(0.1000)	(0.1136)	(0.1647)
2	0.0000	0.3558	0.3566
(not sure)	(0.0968)	(0.1160)	(0.1637)
1	0.1155	0.0279	0.0116
(green)	(0.0499)	(0.0373)	(0.0679)

a low rather than high reward (82% vs. 18%)

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 describes how subjects update beliefs following choices of "blue" ($Y_t = 2$). Therefore, we will not discuss them in great detail.

4 A comparison of methodologies: nonparametric vs. "standard" learning rules

In this section, we compare the predictive ability of our estimated learning rule, vs. alternative learning models which can be calibrated directly from the experimental data. Since the existing literature mainly follows the latter path, this exercise can be considered a comparison of methodologies. In this section, we will refer to our estimated model as the "nonparametric" model, for convenience. We consider two alternative learning rules: Bayesian and reinforcement learning. For each of these cases, the calibration procedure used to fit the model to the data is presented in the appendix.

To begin, we consider Figures 3-5, which contain the raw histograms for the (noisy) measurements of beliefs from the three competing learning models: Figure 3 contains the histogram of the eye tracking measure Z, which is used to pin down beliefs in our estimated learning rules. Figure 4 contains the histogram of the Bayesian posterior probabilities, computed given our experimental design and the observed data. Finally, Figure 5 contains the histogram for the difference in the calibrated valuation measures for the "blue" vs. "green" slot machine, from a TD-leaning reinforcement learning model.

A noteworthy feature is that the histograms for the eye-tracking measure Z and the TD-learning valuations look similar: both are trimodal. The Bayesian posterior mean measure, on the other hand, is unimodal. As we will see later, this implies that the Bayesian learning model tends to predict "smoother" choice behavior than what we observe in the data, whereas both our nonparametric model and the RL model are better at matching the "jumpiness" in the data.

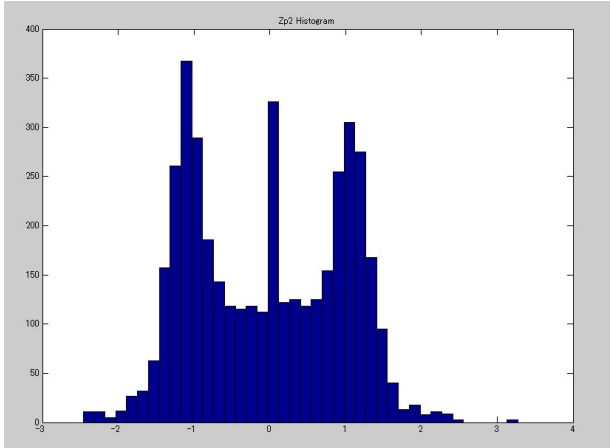

Figure 3: histogram of Z = Zb-Zg

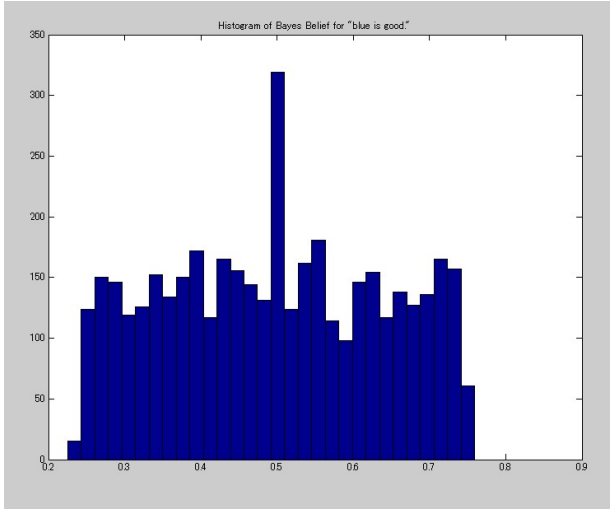

Figure 4: histogram of Bayesian Belief

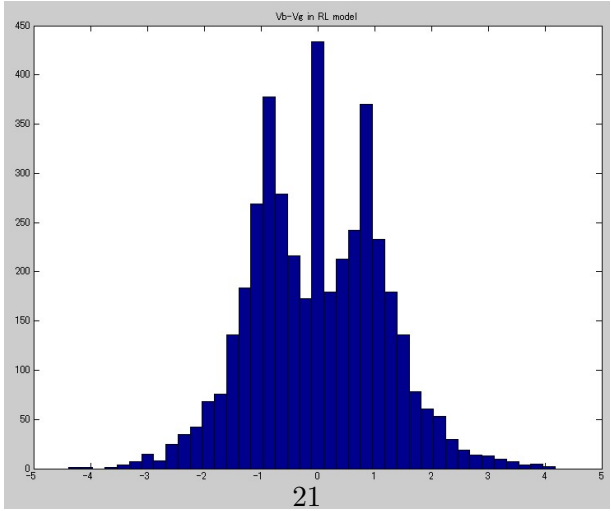

Figure 5: histogram of Vb-Vg in RL

Table 6: Summary statistics for the three models

Panel 1:

X^*	1(green)	2(not sure)	3(blue)
	1988	253	1959

Panel 2:

	mean	median	std.	1/3 quantile	2/3 quantile
B^* (Bayesian Belief)	0.4960	0.5000	0.1433	0.4201	0.5644
$V^*(= V_b - V_g)$	-0.0035	0	1.1152	-0.6588	0.6068

Panel 3: Correlations in the three models

Corr.(X^*, B^*)	0.5789
Corr.(X^*, V^*)	0.5352
Corr.(B^*, V^*)	0.8271

Panel 4: Correlations with observed choices Y (all samples)

Corr.(Y, X^*)	0.9606
Corr.(Y, B^*)	0.5175
Corr.(Y, V^*)	0.5560

Panel 5: Correlations with choices Y (excluding intermediate beliefs)

Corr.(Y, X^*)	0.9909	(keep only X^* =1,3)
Corr.(Y, B^*)	0.6252	(keep only $B^* \notin [1/3 \text{ quant.}, 2/3 \text{ quant.}]$)
Corr.(Y, V^*)	0.6786	(keep only $V^* \notin [1/3 \text{ quant.}, 2/3 \text{ quant.}]$)

Panel 6: Correlations with choices Y (last 10 rounds, first 5 rounds)

	last 10	first 5
Corr.(Y, X^*)	0.9788	0.8665
Corr.(Y, B^*)	0.5267	0.4678
Corr.(Y, V^*)	0.5582	0.5201

Panel 7: Number of exploration choices Y (excluding intermediate beliefs)

Nonparametric	18	(0.46%)
Bayesian	543	(38.8%)
Reinforcement Learning	455	(32.5%)

Panel 8: Correlations with noisy measure Z (NB: Corr.$(Z, Y) = 0.7738$)

Corr.(Z, X^*)	0.7821
Corr.(Z, B^*)	0.4296
Corr.(Z, V^*)	0.4717

Overall summary statistics In Table 6, we present some summary statistics which describe the predictive success of our nonparametric learning model (as given by the optimally-fitted beliefs X_t^*), vs. the Bayesian beliefs B^* and the valuations V^* in the RL 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 RL model, and the Bayesian model. This table contains contains eight panels.

Panel 1 gives the total tally, across all subjects, rounds, and trials, of the number of times the nonparametric beliefs X^* took each of the three values. We see that subjects' beliefs tended to favor green and blue roughly equally, with "not sure" lagging far behind. The almost-equal split 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 round. The second panel shows analogous statistics for the beliefs from the RL and Bayesian models. The RL valuation measure V^* appears largely symmetric and centered around zero, while the average Bayesian B^* lies just below 0.5, thus showing a very slight bias towards green.

Panels 3 and 4 are the key panels in this table. Panels 3 contains the pairwise correlation among (X^*, V^*, B^*), the beliefs from the three models. Obviously, the high correlation (0.8271) between B^* and V^* indicates that, informationally, the beliefs from the Bayesian and RL models are very similar. However, the correlations between our nonparametric beliefs X^* and either B^* and V^* are markedly lower at, respectively, 0.58 and 0.54. Accordingly, the next panel shows that the correlation of X^* with the observed choices Y is much higher (0.9606) than the correlation of choices with the other measures. This is clear evidence about the superior performance of our nonparametric beliefs in predicting subjects' choices.

The next two panels break down the correlation between the observed choices and the difference measures of beliefs, for subsamples of the data. Panel 5 only considers subjects' choices when the implied beliefs are strong (in the sense of taking extreme values). For the nonparametric model, we omitted observations when X^* was estimated to be "not sure", while for the other two models, we omitted observations when beliefs lay between the 1/3 and 2/3 quantile. The results show that when beliefs are strong, the nonparametric model predicts choices almost perfectly (the correlation is 0.99), while the Bayesian and RL beliefs still lag far behind (with correlations of, respectively, 0.63 and 0.68).

In the context of the "explorative/exploitative" dichotomy of choices described earlier,

this implies that our nonparametric model should classify substantially fewer choices as "exploratory" ones (where exploratory behavior is generally defined as making contrarian choices in the face of strong beliefs). This intuition is confirmed in Panel 7, which shows that the nonparametric model classifies only 18 (0.46%) of the subjects' choices as exploratory, while the other two models classify more than twenty-five times those observations as exploratory. This suggests that, at the very least, studies of dynamic choice behavior based on the Bayesian or RL paradigm may seriously misjudge the extent to which subjects engage in exploratory behavior. On the other hand, because exploration is usually a necessary condition of optimal long-run decision-making in probabilistic reversal learning models, our nonparametric learning rules may be substantially less optimal than Bayesian or RL learning rules.

Panel 6 similarly shows that predicted choice behavior using only the last ten rounds of each subjects' data, or the first five rounds, is more accurate for all three models; but as before, the nonparametric model is substantially more accurate. Finally, the bottom panel shows the sample correlation between the eye-movement measure, and the implied beliefs. Not surprisingly, the correlation is much higher for the nonparametric beliefs X^* (since identification of the nonparametric model relies on the monotonicity condition in Assumption 3). The Bayesian and RL beliefs, which do not require Z to compute, exhibit a smaller correlation with Z.

A closer look at individual rounds To gain more insight into the predictive differences between the three models, we plot, in Figures 6-9, the actual choices, as well as subjects' beliefs regarding which slot is better, from the three learning models, for four representative subject-rounds of choices. The actual choices are plotted in crosses (+'s), with higher crosses signifying "blue" and lower crosses signifying "green". The subject's beliefs from the three models are plotted in three different lines.

Figure 6, for trial #4 of subject #6, is typical. As here, subjects typically begin a trial by alternating between "blue" and "green"; the initial "experimentation" period lasted about eight choices, for this particular subject and round. Subsequently, up to the end of the round, choices tend to exhibit more persistence; in this case, subject #6 settled on "blue" for three choices, then "green" for five choices, etc.

Comparing the predicted choices, we see that, generally, all three models perform reasonably well. However, our nonparametric model performs noticeably better, predicting almost all the choices after the initial experimentation period. At the same time, both the Bayesian

and TD-learning model posit "smoother" behavior: for example, after choice #17 (which was "green"), the subject received a low reward. This triggers both the Bayesian and RL beliefs to move "in the direction" of "blue", but not enough to predict a choice of blue in period #18. However, our nonparametric model predicted that choice. This may suggest that human subjects over-attribute negative rewards to structural change in the state S_t, rather than pure random chance.

It is also interesting how differently three models "rationalize" the initial experimentation period. Both the TD-leaning and Bayesian model jumps around a lot in this period, because at the beginning of each round, beliefs and valuations are very sensitive to the rewards. However, the predicted choices for these two models are almost diametrically opposite from the actual observed choices. On the other hand, our nonparametric model explains these initial choices by subjects having beliefs which are "not sure".

Figure 6 as well as the choice probabilities (Table 4) and learning rules (Table 5) are consistent with the notion that humans utilize a "two-mode" algorithm for decision-making under uncertainty, which consists of the two modes of either "searching" or "exploiting". In the initial rounds, subjects utilized random choices to find a "better" slot machine. After several choices, however, once they determine the better one, their choice behavior suddenly changes from random choices to an "exploitative" strategy whereby they keep choosing the preferred one (around 99% in our case), while maintaining a slight exploration probability (around 1%). The change is not gradual, but sudden, as our estimates in Table 4 shows that the choice probabilities are well captured by the three-step choice function, exploiting from green, searching randomly and exploiting from blue.

Figure 7, which shows subject (#4) and round (#6), is similar to the previous figure. However, it is noteworthy that the Bayesian model "misses" the final run of "green" choices. On the other hand, the nonparametric beliefs do a slightly better job of explaining the initial choices, which was also apparent from the summary statistics presented earlier (Panel 6 in Table 6).

Figures 8 and 9, contain instances of choices which were considered "exploratory" (contrarian) through the lens of the Bayesian or RL model, but are predicted by the nonparametric learning model. This happens at choice #9 in Figure 8, and choice #10 in Figure 9. In the first case, we see that the subject's choice of "blue" at trial #8 led to a low reward. This caused the nonparametric belief to update immediately to "green", as suggested by the "extremal" aspect of the belief updating process from the left-hand columns of the learning

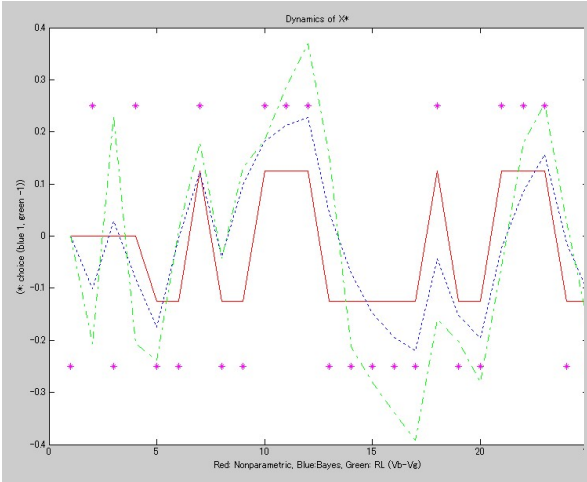

Figure 6: Subject 6, trial 4

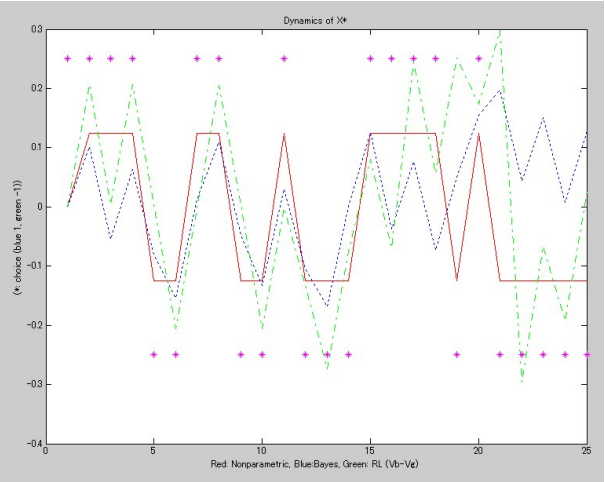

Figure 7: Subject 4, trial 6

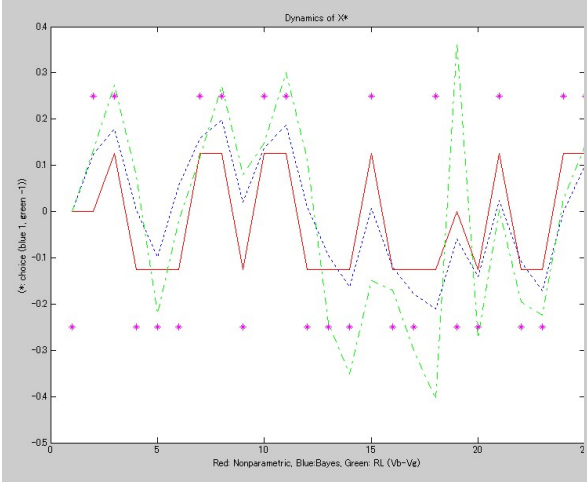

Figure 8: Subject 5, trial 8

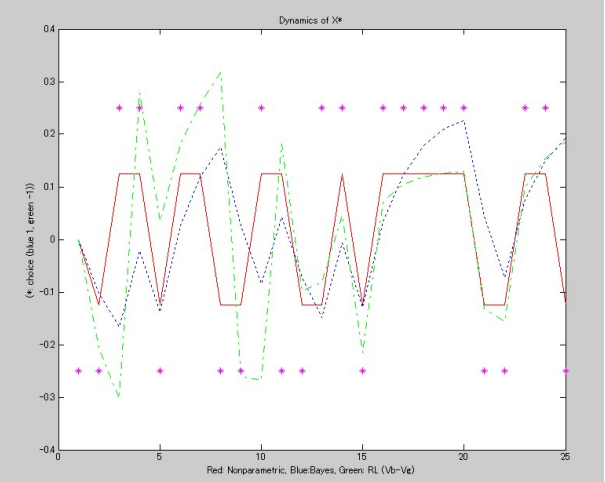

Figure 9: Subject 1, trial 3

rule estimates in Table 5. The problem with the Bayesian and RL models here is that they update too smoothly, and cannot capture the jumpiness in the choice behavior.

Choice #10 in Figure 9 presents an even more striking case, as this choice of "blue" was preceeded by a choice of "green" accompanied by a high reward. Here, both the Bayesian and RL models update strongly towards "blue", but the nonparametric belief is able to explain the surprising choice of "green". This can be attributed to the estimates in the second panel of Table 5, which show that even after a profitable choice of "green", beliefs may still update, with 11% probability, to "blue". This possibility appears to be missed in the Bayesian and RL models.

On the other hand, choice #18 in Figure 8, in which the subject jumped immediately back to "blue" after receiving positive rewards from her two immediately preceding choices of "green" at trials #16 and #17, was completely unanticipated, and is classified as an explorative choice by all three models. Hence, while the nonparametric model seems to accommodate subjects' jumpy behavior in the experiment better than the Bayesian or RL models, the actual behavior is still more haphazard than what the nonparametric model allows for.

By itself, the superior performance of the nonparametric model relative to the other two models, which we have documented here, is not surprising, because the nonparametric model was actually estimated from the observed choices. However, our nonparametric approach relies crucially on the validity of the auxiliary measure Z_t as a measure of beliefs (as encapsulated in our Assumption 3). Therefore, if eye movements were an unreliable measure of beliefs, then our entire approach would fail, and should produce learning rules with little predictive value for subjects' choices.[5] Clearly, as we have shown here, this has not been the case. As a reuslt, it suggests that nonparametric estimation of subjects' learning rules, as an approach for assessing learning in experimental settings, may be a more convenient and better approach than fitting pre-specified learning models to the data, which has been the prevalent practice in most of the literature.

5 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 mea-

[5] Of course, by construction, the nonparametric beliefs always predict the auxiliary measure Z_t well.

surements 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, non-smooth updating following positive "exploitative" choices – which fit the observed choice data better than alternative parameterized learning rules which are commonly assumed in the literature, including Bayesian and reinforcement learning rules.

Our nonparametric estimator for subjects' choice probabilities and learning rules is easy to implement. Potentially, it can also be applied to other experimental settings where auxiliary measures of subjects' beliefs and valuations are available, such as the typical neuroscience fMRI setting.

References

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

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

Behrens, T., M. Woolrich, M. Walton, and M. Rushworth (2007): "Learning the value of information in an uncertain world," *Nature Neuroscience*, 10(9), 1214–1221.

Chan, T. Y., and B. H. 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.

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

Daw, N., J. O'Doherty, P. Dayan, B. Seymour, and R. Dolan (2006): "Cortical substrates for exploratory decisions in humans," *Nature*, 441(7095), 876–879.

Doya, K. (2002): "Metalearning and neuromodulation," *Neural Networks*, 15(4-6), 495–506.

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.

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

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

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.

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

Hu, Y., and M. Shum (2008): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," Jonhs Hopkins University, Dept. of Economics working paper #543.

Imai, S., N. Jain, and A. Ching (2009): "Bayesian Estimation of Dynamic Discrete Choice Models," *Econometrica*, 77, 1865–1899.

Krajbich, I., C. Armel, and A. Rangel (2007): "Visual attention drives the computation of value in goal–directed choice," Discussion paper, Working Paper, Caltech.

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

Odean, T., M. Strahilevitz, and B. Barber (2004): "Once Burned, Twice Shy: How Naive Learning and Counterfactuals Affect the Repurchase of Stocks Previously Sold," Discussion paper, mimeo., UC Berkeley, Haas School.

Pakes, A., and P. McGuire (2001): "Stochastic Algorithms, Symmetric Markov Perfect Equilibrium, and the 'Curse' of Dimensionality," *Econometrica*, 69, 1261–1282.

Payzan, É., and P. Bossaerts (2009): "Decision-making under uncertainty in dynamic settings: an experimental study," .

Rangel, A. (2008): "The computation and comparison of value in goal-directed choice," *Neuroeconomics: Decision-making and the brain. P. Glimcher, C. Camerer, E. Fehr, & R. Poldrack (eds). New York: Elsevier.*

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

Samejima, K., K. Doya, Y. Ueda, and M. Kimura (2004): "Estimating internal variables and parameters of a learning agent by a particle filter," *Advances in Neural Information Processing Systems*, 16.

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

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

Yoshida, W., and S. Ishii (2006): "Resolution of uncertainty in prefrontal cortex," *Neuron*, 50(5), 781–789.

6 Appendix: Additional details on computation of nonparametric, Bayesian, and RL learning rules

In section 4, 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.

6.1 Belief dynamics X^* in our 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 $P(\{X_t^*\} \,|\, \{Y_t, Z_t, R_t\})$, the conditional ("posterior") probability of the beliefs, given the observed sequences of choices, eye-movements, and rewards. Because

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

where the second term on the RHS of the equation above does not depend on X_t^*, it is equivalent to maximize $P(\{X_t^*, Y_t, Z_t\} \,|\, \{R_t\})$ with respect to $\{X_t^*\}$.

Because of the Markov structure, the joint log-likelihood of $\{Y_t, Z_t, X_t^*\}_{t=1}^{25}$ is:

$$\begin{aligned} &\log L(\{Y_t, Z_t, X_t^*\}|\{R_t\}) \\ &= \sum_{t=1}^{24} \log(P(Y_t|X_t^*)P(Z_t|X_t^*)P(X_{t+1}^*|X_t^*, R_t, Y_t)) + \log(P(Y_{25}|X_{25}^*))P(Z_{25}|X_{25}^*). \end{aligned} \tag{6}$$

We plug in our nonparametric estimates of $P(Y|X^*)$, $P(Z|X^*)$ 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 round). To facilitate this optimization problem, we derive the optimal sequence of beliefs using a dynamic-programming (Viterbi) algorithm; cf. Ghahramani (2001).

6.2 Bayesian Learning Model

A Bayesian learner uses Bayes rule to update her beliefs. Let B_t^* denote the belief, or prior probability, that the blue slot machine is the good one at the start of the trial t. After her choice, she observes reward R_t. Let ${B'}_t^*$ denote the posterior belief, the probability that the blue slot machine is the good one after R_t is observed. The posterior probability is derived using Bayes rule:

$$ {B'}_t^* = \frac{P(R_t|S_t = 1) \cdot B_t^*}{P(R_t|S_t = 1) \cdot B_t^* + P(R_t|S_t = 2) \cdot (1 - B_t^*)} \tag{7} $$

At the end of each trial, the state S_t may change with 15% chance. The Bayesian learner takes this into account, so that the prior probability on "blue" at the start of trial $t + 1$ is the posterior probabilities weighted by the state transition probabilities:

$$ B_{t+1}^* = P(S_{t+1} = 1|S_t = 1) \cdot {B'}_t^* + P(S_{t+1} = 1|S_t = 2) \cdot (1 - {B'}_t^*). \tag{8} $$

In this way, given the initial beliefs $B_1 = 0.5$, we can use Eqs. (7) and (8) to compute the sequence of Bayesian beliefs, $\{B_t^*\}$, corresponding to the observed sequences of choices and rewards $\{Y_t, R_t\}$.

6.3 Reinforcement Learning Model

Here we use a variant of the TD (Temporal-Difference)-Learning models (Sutton and Barto (1998), section 6). The value of an action is learned by the reward that is expected after taking that action. Let $V_{b(g)}^t$ denote the "current" (ie. beginning of trial t) action value function for the blue (green) slot machine. The value updating rule for a "One-step TD-Learning" model is defined as:

$$ V_{c_t}^{t+1} \longleftarrow V_{c_t}^t + \alpha\delta_t. \tag{9} $$

where $c_t \in \{b, g\}$ is the choice taken in trial t, α denotes the learning rate, and δ_t denotes the "prediction error" for trial t (defined below). For greater model flexibility, we allow the parameter α to take two values, for positive reward and negative reward. For instance, if

$c_t = b$ (so "blue" was chosen in trial t), then the TD learning rule implies that V_b is updated by an amount equal to the prediction error δ_t, weighted by the learning parameter α (with larger values of α indicating an increased sensitivity to the outcome of trial t).

The prediction error δ_t is equal to

$$\delta_t = (R_t + \gamma E[V^t_{c_{t+1}}|t]) - V^t_{c_t} \tag{10}$$

the difference between $(R_t + \gamma E[V^t_{c_{t+1}}|t])$ (the observed reward in trial t plus the discounted expected value from the next trial), and $V^t_{c_t}$ (the current expected valuation). We assume the discount factor $\gamma = 0.9$.

The variant of TD-Learning (SARSA) used here (Sutton and Barto (1998), p. 149) computes the expected value function $E[V_{c_{t+1}}|t]$ using the current choice probabilities of choosing action $c(t+1)$. Let P^t_c denote the current probability of choosing c. We adopt the conventional "softmax" (ie. logit) choice probability function with the inverse temperature parameter β:

$$P^t_{c_t} = \frac{e^{\beta V^t_{c_t}}}{\sum_{c'_t} e^{\beta V^t_{c'_t}}} \tag{11}$$

With this functional form for the choice probabilities, the expected value function from trial $t+1$ is computed as,

$$E[V_{c_{t+1}}|t] = \sum_{c'_{t+1} \in (b,g)} P^t_{c'_{t+1}} V^t_{c'_{t+1}}. \tag{12}$$

The choice function (Eq. (11)) can be rewritten as a function of the difference $V^*_t \equiv V^t_b - V^t_g$. The current choice probability for the blue slot machine is,

$$P^t_b = \frac{e^{\beta(V^t_b - V^t_g)}}{1 + e^{\beta(V^t_b - V^t_g)}} = \frac{e^{\beta V^*_t}}{1 + e^{\beta V^*_t}}. \tag{13}$$

To obtain estimates for β and two α, we apply maximum likelihood estimation. The estimates we obtained from the data were:

$$\begin{aligned} &\beta = 0.7584 \\ &\alpha \text{ for large reward } (R_t = 2) = 1.6531 \\ &\alpha \text{ for small reward } (R_t = 1) = 1.0552. \end{aligned} \tag{14}$$

We plug in these values into the TD-Learning model to derive a sequence of $\{V_t^* \equiv V_b^t - V_g^t\}$, which are analogous to the belief measures from the nonparametric and Bayesian learning models.

7 Appendix: Some additional details on discretization

In what follows, we use Z_{pt} to denote the continuous-valued eye-tracking measure, and Z_t the discretized version. For the two-value discretization, we discretize as follows:

$$Z_t = \begin{cases} 1 & \text{if } Z_{pt} < 0 \\ 2 & \text{if } Z_{pt} \geq 0 \end{cases}$$

As we discussed before, since we do not find any color bias toward blue nor green, discretization of Z_{pt} around 0 should be reasonable. The sample size for $Z_t = 1$ (green) is 2032 and that for $Z_t = 2$ (blue) is 2168 (Table 2). One might worry that the classification for $Z_{pt} = 0$ observations in blue yields a certain bias. Actually we observe 230 observations with $Z_{pt} = 0$ and classify them as $Z_t = 2$ (blue) for convenience. However, it turns out to matter little whether these 230 observations are classified as blue or green.

For the three-value discretization, we discretize Z_{pt} as follows:

$$Z_t = \begin{cases} 1 & \text{if } Z_{pt} < \text{-sid.} \\ 2 & \text{if -sid.} \leq Z_{pt} \leq \text{sid.} \\ 3 & \text{if sid.} < Z_{pt} \end{cases}$$

where "sid." is the factor used to normalize Z_{pt}. The choice of the value for sid. does not affect the estimation results essentially. Figure 3 is the histogram of Z_{pt}. The shape of the histogram appears to have a mixture of three different distributions, one large distribution centered around -1, another large distribution centered around 1 and the other relatively small distribution centered around 0. As the baseline, we set sid. to 0.20 to normalize the three distributions effectively. However, we do not find any difference in the estimation results either qualitatively nor significantly even if we change the sid. parameter from 0.05 to around 0.40, except for the measurement error probabilities $P(Z_t|X^*)$, suggesting that the model is robust for different classifications. This robustness is reasonable since our nonparametric model does not assume any functional form for measurement error in Z_{pt}. The last panel in Table 2 shows the sample sizes for the each classifications, 1887 for $Z_t = 1$, 540 for $Z_t = 2$ and 1773 for $Z_t = 2$ in the baseline ($sid. = 0.20$).

Table 7: Correlations (Y, Z_p) in each classification (All sample = 0.7647)

sid. = 0.05	Size	Correlation
$Z_t = 1$ (green)	2015	0.3223
2 (not sure)	255	-0.0599
3 (blue)	1930	0.2346

sid. = 0.20 (baseline)	Size	Correlation
$Z_t = 1$ (green)	1887	0.2845
2 (not sure)	540	0.2156
3 (blue)	1773	0.1706

sid. = 0.40	Size	Correlation
$Z_t = 1$ (green)	1725	0.1462
2 (not sure)	869	0.2777
3 (blue)	1606	0.0991

At this point, one might ask whether the beliefs, as captured by the unobserved variable X_t^*, should take more than three values. What is the appropriate number of discretization for the belief space? Next, we present some evidence suggesting that a three-value discretization for X_t^* captures the observed choice behavior well. Since X_t^* is unobserved, we examine the form of the choice probability $P(Y_t|Z_t)$ conditional on the observed measure Z_t instead of the unobserved X_t^*. From our estimates of the measurement probabilities, we know that Z_t tracks X_t^* rather closely, so that $P(Y_t|Z_{pt})$ should not be too different from $P(Y_t|X_t^*)$. If $P(Y_t|X_t^*)$ is a step function with three steps, as we have assumed in the "three-value" specification, then also $P(Y_t|Z_{pt})$ should be a similar function.

Table 7 shows the correlations between Y and Z_p, broken up into the three ranges of Z_p corresponding to the three discretized values $Z \in \{1, 2, 3\}$, and also for three different values of the standard deviation σ parameter. Although the correlation in the whole sample is 0.7647, the correlations within each of the three ranges of Z_p drop significantly, ranging from even negative values to values around 0.30. This suggests that the discretization of Z_t is sufficient, because after the discretization, most of the predictability in choices is *across* the different discretized values of Z, rather than within these values.

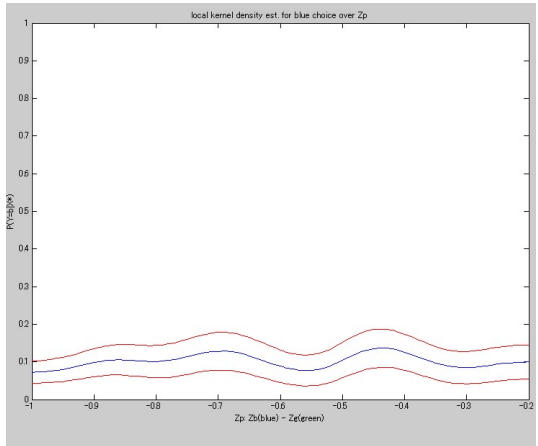

Figure 10: Local kernel estimation for choice prob. over Zp (Z = 1, $Zp < -sid.$), bandwidth=$0.2\times\sigma^2$

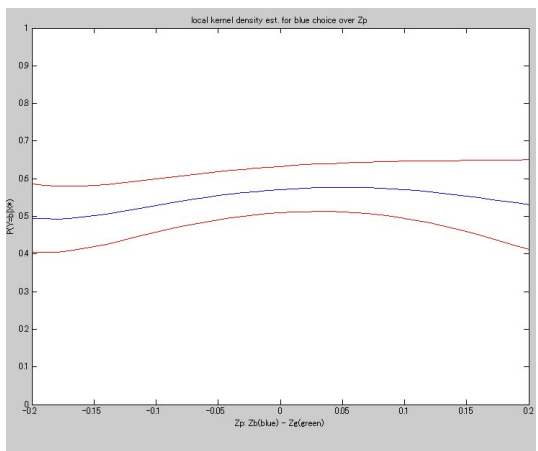

Figure 11: Local kernel estimation for choice prob. over Zp (Z = 2, $-sid. \leq Zp \leq sid.$), bandwidth=$0.2\times\sigma^2$

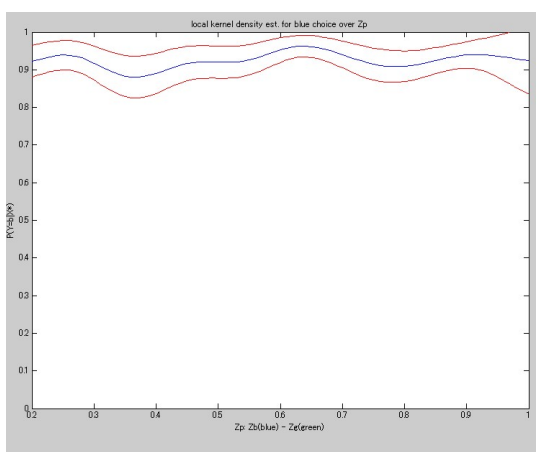

Figure 12: Local kernel estimation for choice prob. over Zp (Z = 3, $sid < Zp$), bandwidth=$0.2\times\sigma^2$

Examining this further, we performed a kernel regression of Y on Z_p, separately for the three different ranges of the Z_p variable. We consider a Gaussian kernel with bandwidth $h = 0.2 \times \sigma^2$. The three graphs in Figures 10-12 show the results for the kernel regression, separately for the ranges of Z_p corresponding to the discretized values $Z = 1, 2, 3$. We see that the kernel estimates of $P(Y|Z)$ are practically constant in each graph. Figure 10 indicates that for low values of Z_p, corresponding to $Z = 1$, the probability of choosing the blue slot machine is around 7.5%. In Figure 11, for the range in Z_p corresponding to $Z = 2$, the probability of choosing blue is around 52.5%, while when $Z = 3$, corresponding to the largest values of Z_p, the probability of choosing blue is around 95% (as seen in Figure 10). These graphs support the earlier finding that a three-value discretization of Z_p is sufficient to capture most of the variation on Y, which is why we focus on discrete specifications in our econometric models.