frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 04 March 2014
doi: 10.3389/fncom.2014.00018

Bayesian deterministic decision making: a normative account of the operant matching law and heavy-tailed reward history dependency of choices

Hiroshi Saito*[1], *Kentaro Katahira*[2,3,4], *Kazuo Okanoya*[3,4,5] *and Masato Okada*[1,3,4]

[1] *Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan*
[2] *Center for Evolutionary Cognitive Sciences, The University of Tokyo, Tokyo, Japan*
[3] *RIKEN Brain Science Institute, Wako, Japan*
[4] *Okanoya Emotional Information Project, Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency, Wako, Japan*
[5] *Department of Life Sciences, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan*

Edited by:
Stefano Fusi, Columbia University, USA

Reviewed by:
Maneesh Sahani, University College London, UK
Emili Balaguer-Ballester, Bournemouth University, UK
Brian Lau, Centre de Recherche de l'Institut du Cerveau et de la Moelle Epinière, France

****Correspondence:***
Masato Okada, Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, Kashiwa-campus of the University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba 277-8561, Japan
e-mail: okada@k.u-tokyo.ac.jp

The decision making behaviors of humans and animals adapt and then satisfy an "operant matching law" in certain type of tasks. This was first pointed out by Herrnstein in his foraging experiments on pigeons. The matching law has been one landmark for elucidating the underlying processes of decision making and its learning in the brain. An interesting question is whether decisions are made deterministically or probabilistically. Conventional learning models of the matching law are based on the latter idea; they assume that subjects learn choice probabilities of respective alternatives and decide stochastically with the probabilities. However, it is unknown whether the matching law can be accounted for by a deterministic strategy or not. To answer this question, we propose several deterministic Bayesian decision making models that have certain incorrect beliefs about an environment. We claim that a simple model produces behavior satisfying the matching law in static settings of a foraging task but not in dynamic settings. We found that the model that has a belief that the environment is volatile works well in the dynamic foraging task and exhibits undermatching, which is a slight deviation from the matching law observed in many experiments. This model also demonstrates the double-exponential reward history dependency of a choice and a heavier-tailed run-length distribution, as has recently been reported in experiments on monkeys.

Keywords: decision making, operant matching law, Bayesian inference, dynamic foraging task, heavy-tailed reward history dependency

1. INTRODUCTION

Does the brain play dice? This is a controversial question about the underlying processes of the brain in making a choice from several alternatives: Does the brain decide deterministically with some internal decision variables? Or does it calculate the probability of choosing individual alternatives and cast a "biased die" (Sugrue et al., 2005)? The former strategy is suggested according to our everyday experience. However, it is possible to think that choices emerge probabilistically by observing a sequence of decisions in a repetitive task. Herrnstein conducted a foraging experiment where a pigeon was placed into a box that was equipped with two keys and when a key was pressed it was rewarded with concurrent variable-interval schedules. He found a relationship between rewards and choices known as the "operant matching law" (Herrnstein, 1961). The law states that the fraction of the number of times one alternative is chosen against the total number of choices matches the fraction of the cumulative reward obtained from the alternative against the total reward. Behaviors satisfying the law have been observed in a variety of task paradigms and across species (de Villiers and Herrnstein, 1976; Gallistel, 1994; Anderson et al., 2002). Several learning models have been proposed to account for matching behavior (Corrado et al., 2005; Lau and Glimcher, 2005; Loewenstein and Seung, 2006; Soltani and Wang, 2006; Sakai and Fukai, 2008a; Simen and Cohen, 2009). These models have a commonality in that a model learns the probabilities of choosing each alternative directly, and then a choice is made stochastically. However, it is yet unknown whether matching behaviors can be accounted for by a deterministic model.

Here, we propose deterministic Bayesian decision making models for a two-alternative choice task. Our models stand on the incorrect but conceivable postulate that animals have a belief that the choice made in one trial does not affect a reward in subsequent trials. The models estimate the unknown reward probabilities for each alternative and deterministically choose the alternative that has the highest reward probability according to the *winner-take-all* principle. We first study a model with belief that the environment does not change. Note that this is an extension of the fixed belief model (FBM) (Yu and Cohen, 2009) for the two-alternative choice task. We demonstrate that this model satisfies the matching law in a steady state in static foraging tasks, in which reward baiting probabilities are fixed, but not in dynamic foraging tasks, in which the reward baiting probabilities change abruptly. Then, we devise two models that forget past experience

and exhibit matching behaviors even in dynamic tasks. Moreover, these models can explain *undermatching*, which is a phenomenon observed across different species (Baum, 1974; de Villiers and Herrnstein, 1976; Baum, 1979; Gallistel, 1994; Anderson et al., 2002; Sugrue et al., 2004; Lau and Glimcher, 2005). We test these models by comparing their predicted reward history dependencies and run-length distributions to those seen in a monkey experiment.

2. RESULTS

We studied deterministic Bayesian decision making models that demonstrated matching behaviors in a foraging task. The foraging task is a decision making task that simulates a foraging environment where an animal chooses one out of several foraging alternatives. There are two alternatives in this study although our results do not depend on this. We employed discrete trial-to-trial tasks that have often been used in recent experiments (Sugrue et al., 2004; Corrado et al., 2005; Lau and Glimcher, 2005). Each alternative has binary baiting state f_i ($i \in \{1, 2\}$ is the index of an alternative), where $f_i = 1$ if a reward is baited and $f_i = 0$ otherwise. If $f_i = 0$, a reward is baited ($f_i = 1$) at the beginning of each trial by baiting probability λ_i^t, where t represents the number of the trial. If the baiting probabilities are fixed across trials, the task is called a *static* foraging task, otherwise it is called a *dynamic* foraging task (Sugrue et al., 2004). Suppose that r_i^t indicates whether a subject receives a reward ($r_i^t = 1$) or not ($r_i^t = 0$), and c_i^t indicates whether the subject chooses alternative i ($c_i^t = 1$) or not ($c_i^t = 0$) in trial t. When the subject chooses a baited alternative, i.e., $f_i = 1$ and $c_i^t = 1$, the baited reward is consumed ($f_i \leftarrow 0$). This reward schedule is known as a "concurrent variable-interval schedule"(Baum and Rachlin, 1969).

Whichever alternative the subject chooses in the foraging task, the choice can affect the reward probabilities of alternatives in the future. Therefore, the optimal strategy is not to exclusively choose the foraging alternative that has the highest baiting probability. A behavioral strategy obeying the matching law is known to be nearly optimal for this task (Baum, 1981). Formally, the law states that

$$\frac{\bar{R}_i^t}{\sum_j \bar{R}_j^t} = \frac{\bar{C}_i^t}{\sum_j \bar{C}_j^t}, \tag{1}$$

where $\bar{R}_i^t$ and $\bar{C}_i^t$ correspond to the total reward obtained from alternative i and the number of choices of alternative i until trial t. It is known that human and animal behaviors in these kinds of tasks are well described by the generalized matching law (Baum, 1974)

$$\log(\bar{R}_1^t/\bar{R}_2^t) = s\log(\bar{C}_1^t/\bar{C}_2^t) + \log k, \tag{2}$$

where s is sensitivity and k is bias. Equation (2) is equivalent to (1) if both s and k are unities.

2.1. SIMPLE BERNOULLI ESTIMATORS

First, we studied a simple normative Bayesian decision making model to clarify the underlying feasible computation for matching behaviors. Suppose that a subject makes a decision simply depending on its estimates of the reward probabilities for the alternatives. The estimate can be formally described as

$$P_i^{t+1} = p(r_i^{t+1} = 1 | R^t, C^t), \tag{3}$$

where R^t is a list of reward vectors $\boldsymbol{r}^t = (r_1^t, r_2^t)$ from trials 1 to t and C^t is a list of choice vectors $\boldsymbol{c}^t = (c_1^t, c_2^t)$ from trials 1 to t. The model employs a *winner-take-all* (WTA) strategy, i.e., it chooses the alternative that has the highest P_i^t. The model requires an assumption about a reward assignment mechanism to estimate P_i^{t+1}. One simple and conceivable assumption is that a choice is rewarded according to hidden reward probability μ_i^t that is irrelevant to the past reward and choice history, i.e., $p(r_i^t = 1) = \mu_i^t$. This assumption is incorrect for our tasks but we have assumed that the model employs it and predicts μ_i^t by Bayesian inference. Hence, P_i^{t+1} is given by the predictive distribution over μ_i^t:

$$P_i^{t+1} = \int_0^1 d\mu\, \mu\, p(\mu_i^{t+1} = \mu | R^t, C^t). \tag{4}$$

Note that $p(\mu_i^{t+1} = \mu | R^t, C^t)$ can include a model's belief about the change of μ_i^t in between trials. Our first model assumes that μ_i^t is time invariant, i.e., $p(\mu_i^{t+1} = \mu | R^t, C^t) = p(\mu_i^t = \mu | R^t, C^t)$. The posterior distribution for an alternative is not updated if the alternative is not chosen. If it is chosen, the posterior distribution is updated

$$\begin{aligned} p(\mu_i^t = \mu | R^t, C^t) &\propto p(r_i^t | \mu_i^t = \mu) p(\mu_i^{t-1} = \mu | R^{t-1}, C^{t-1}) \\ &= \mu^{r_i^t}(1-\mu)^{1-r_i^t} p(\mu_i^{t-1} = \mu | R^{t-1}, C^{t-1}). \end{aligned} \tag{5}$$

We employ the Beta prior, $p(\mu_i^0 = \mu) = \mathrm{Beta}(\mu | a, b)$, which is a conjugate for the likelihood. Note that we set the hyperparameters, $a = b = 1$, to make the prior non-informative in all simulations. Therefore, the posterior becomes a Beta distribution:

$$p(\mu_i = \mu | \bar{R}_i^t, \bar{C}_i^t) = \mathrm{Beta}(\mu | \bar{R}_i^t + a, \bar{C}_i^t - \bar{R}_i^t + b). \tag{6}$$

From Equations (4) and (6), we obtain

$$P_i^{t+1} = \frac{\bar{R}_i^t + a}{\bar{C}_i^t + a + b}. \tag{7}$$

This model is a natural extension of FBM (Yu and Cohen, 2009) to the two-alternative choice task (for this reason, we will refer to our model as FBM). An alternative is repeatedly chosen while its predictive distribution is higher than those of the other due to the WTA strategy. Because the empirical probability of reward for an alternative converges to its baiting probability in repeated choices, P_i^t gradually approaches to λ_i and the variance of P_i^t decreases. As a result, FBM tends to choose exclusively the high payoff alternative after a large number of observations. Hence, the matching law [Equation (1)] is satisfied in $t \to \infty$ because such a exclusive choice unboundedly increases both $\bar{R}_i^t$ and $\bar{C}_i^t$ of the high payoff alternative.

We simulated FBM in static and dynamic foraging tasks. The time course for the predictive distributions is shown in **Figure 1A**. As can be expected, both predictive distributions approach the respective baiting probabilities and FBM behavior converges to exclusive choice of the high payoff alternative in static foraging tasks. However, the steady-state choice behavior of animals in static concurrent VI schedules has not been thought to be exclusive (Baum, 1982; Davison and McCarthy, 1988; Baum et al., 1999). It might be that there are not enough trials for choice behavior to actually reach a steady state. **Figures 1B,C** plot the log ratios of rewards and choices in both tasks. The marginal histograms indicate the FBM's strong preference for the alternative that has the highest baiting probability, because most pairs of log ratios lie near the endpoints of the matching line. We found that bias is nearly zero and sensitivity is nearly one in the static foraging tasks (**Figure 1B**) by least-square fitting the generalized matching law [Equation (2)] to the data. Therefore, the model exhibits matching behavior in the static foraging tasks. However, the model no longer exhibits matching behavior in dynamic foraging tasks, a result that is inconsistent with the behavior of monkeys (Corrado et al., 2005) (**Figure 1C**). This can be because the model adheres to past experience and cannot adapt rapidly to changes in the environment.

2.2. EXTENDED BERNOULLI ESTIMATORS

One possible way of improving the model to enable it to rapidly adapt to changes in the environment is to introduce a forgetting mechanism for past rewards and choice history. We therefore assume a simple extended model, which utilizes only the L most recent rewards and choices for the estimates. Hence, the predictive distribution becomes

$$P_i^{t+1} = \frac{(\sum_{l=0}^{L-1} r_i^{t-l}) + a}{(\sum_{l=0}^{L-1} c_i^{t-l}) + a + b}. \tag{8}$$

We refer to this model as windowed FBM (WFBM).

Another possibility may be derived from the idea that humans and animals may innately believe their environment is volatile. Here, we propose a model that estimates time-varying reward probabilities. Although there are several ways to model a belief of a volatile environment, we assume our model believes that μ_i^t remains unchanged with probability α, or else (with probability $1-\alpha$) changes completely. This idea is derived from the dynamic belief model (DBM), proposed by Yu and Cohen as a model of sequential effect (Yu and Cohen, 2009). Our model is a natural extension of DBM to a two-alternative choice task. Thus, we refer to our model as DBM. The transition of μ_i^t is modeled as a mixture of the posterior and prior distributions

$$p(\mu_i^{t+1} = \mu | R^t, C^t) = \alpha p(\mu_i^t = \mu | R^t, C^t) + (1-\alpha)\mathrm{Beta}(\mu | a, b), \tag{9}$$

where $0 \leq \alpha \leq 1$ represents the model's expectations of the stability of the environment. However, the posterior distribution is no longer a Beta distribution:

$$\begin{aligned} &p(\mu_i^t = \mu | R^t, C^t) \\ &= p(\mu_i^t = \mu | r_i^t, c_i^t = 1, R^{t-1}, C^{t-1})^{c_i^t} p(\mu_i^t = \mu | R^{t-1}, C^{t-1})^{1-c_i^t} \\ &= \left[\left(\frac{p(r_i^t = 1 | \mu_i^t = \mu)}{p(r_i^t = 1 | R^{t-1}, C^{t-1})} \right)^{r_i^t} \left(\frac{p(r_i^t = 0 | \mu_i^t = \mu)}{p(r_i^t = 0 | R^{t-1}, C^{t-1})} \right)^{1-r_i^t} \right]^{c_i^t} \\ &\quad p(\mu_i^t = \mu | R^{t-1}, C^{t-1}) \\ &= \left[\left(\frac{\mu}{P_i^t} \right)^{r_i^t} \left(\frac{1-\mu}{1-P_i^t} \right)^{1-r_i^t} \right]^{c_i^t} p(\mu_i^t = \mu | R^{t-1}, C^{t-1}), \end{aligned} \tag{10}$$

where we use Equation (3). Then, predictive distribution P_i^t is calculated with Equations (4), (10), and (11). Note that these models are equivalent to FBM when $L \to \infty$ and $\alpha = 1$.

Figure 2 has the time courses for the predictive distributions of WFBM and DBM, and the posterior distributions of DBM in the dynamic foraging task. Neither model is stuck on one alternative and can follow the changes in schedules as expected. There is a clear difference in the predictive distribution trajectories. Because WFBM exploits recent samples, its predictive distribution for the unchosen alternative can approach the true baiting probability. DBM's predictive distribution for the unchosen alternative, on the other hand, is only retracted to the mean of the prior, i.e., 0.5. Both models demonstrate matching behaviors even in the dynamic foraging task (**Figure 3**). More precisely, the behaviors slightly deviate from the matching law toward an unbiased choice. This phenomenon is known as *undermatching* (Baum,

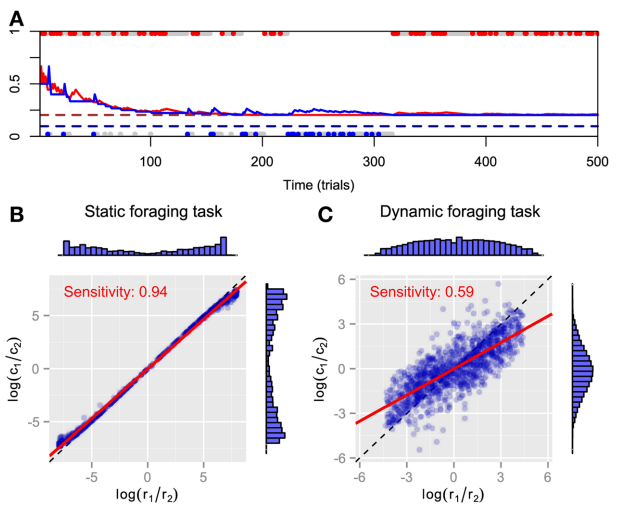

FIGURE 1 | Simulation results for FBM. (A) Time course of predictive distributions for alternatives #1 (red solid line) and #2 (blue solid line) in static foraging task. Dashed lines indicate baiting probabilities of alternatives #1 (dark red) and #2 (dark blue). Upper and lower dots respectively represent choices for alternatives #1 and #2 in that trial and colored dots (red or blue) represent that the model received a reward at that trials. **(B)** Reward log ratios as a function of count log ratios in static and **(C)** dynamic foraging tasks. Blue symbols represent pairs of log ratios calculated in block where baiting probabilities are fixed and distributions of dots are represented by marginal histograms. Red line indicates best-fitted line to points and inner text shows its slope, i.e., sensitivity parameter of generalized matching law. Dashed line is identity line.

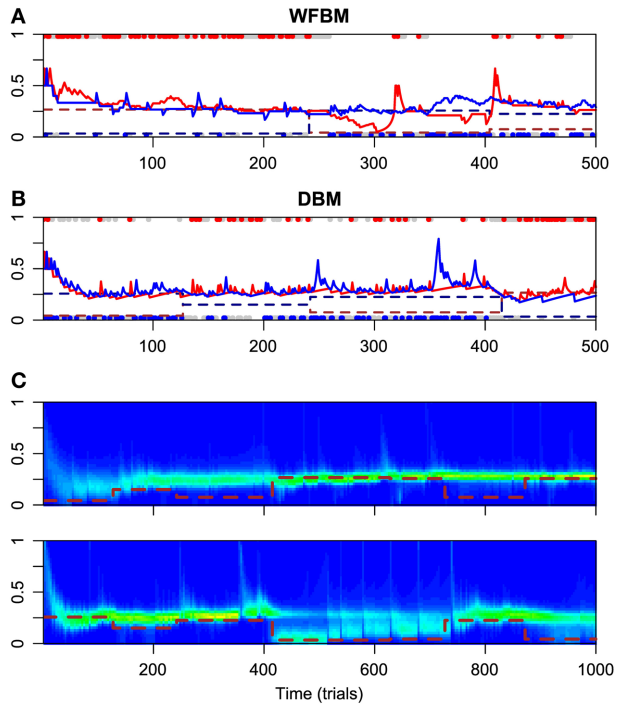

FIGURE 2 | Simulation results for WFBM and DBM in dynamic foraging task. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. **(A)** Time course of predictive distributions of WFBM and **(B)** DBM. Details in figure are described in caption of **Figure 1A**. **(C)** Time course of posterior distributions of DBM for reward probabilities of alternative #1 (top) and #2 (bottom). Brown dashed lines are baiting probabilities for respective alternatives.

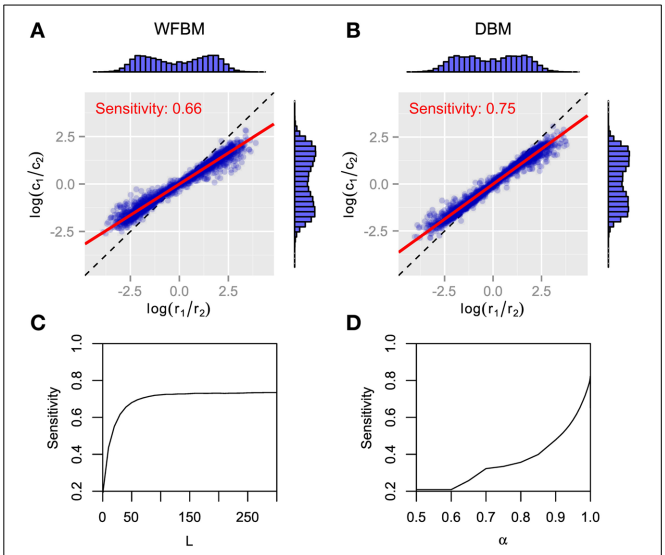

FIGURE 3 | Analytical results for matching behavior of WFBM and DBM in dynamic foraging tasks. (A,B) Reward log ratios as a function of count log ratios. Details in figure are described in caption of **Figures 1B,C**. Simulation parameters were set to $L = 40$ and $\alpha = 0.99$. **(C,D)**Sensitivity as a function of parameters of WFBM and DBM.

1979). Because the models' parameters L and α control the effect of past experience, the degree of undermatching is controlled by the parameters. The sensitivities that were fitted in the experiments were in a range of about 0.44 to 0.91 (Hinson and Staddon, 1983; Corrado et al., 2005; Lau and Glimcher, 2005). Hence, we basically focused on parameter regions $10 \leq L$ and $0.9 \leq \alpha$.

The dependence of choices on reward history has been studied in several monkey experiments. An exponential shaped dependency was first reported (Sugrue et al., 2004) and then heavier-tailed dependencies were reported (Corrado et al., 2005; Lau and Glimcher, 2005). We tested our models by calculating the dependence of choices on reward history (**Figure 4A**). Suppose that dependency is expressed with a linear filter kernel $\kappa(i)$ as in previous studies. The kernel is calculated by minimizing the following Wiener-Hopf equation,

$$\frac{1}{2}\sum_t \left[(c_1^t - c_2^t) - \sum_{i=1}^{K} \kappa(i)(r_1^{t-i} - r_2^{t-i}) \right]^2. \qquad (11)$$

Then, we fit the exponential filter and double-exponential filter that were introduced by Corrado et al. (2005) to the normalized kernel:

$$\epsilon_1(i) = \frac{\exp(-i/\tau_0)}{\sum_{k=1}^{K}\exp(-k/\tau_0)},$$
$$\epsilon_2(i) = \rho\,\frac{\exp(-i/\tau_1)}{\sum_{k=1}^{K}\exp(-k/\tau_1)} + (1-\rho)\frac{\exp(-i/\tau_2)}{\sum_{k=1}^{K}\exp(-k/\tau_2)}, \qquad (12)$$

where τ_0 and $\tau_1 \leq \tau_2$ are time constants and $0 < \rho < 1$ is the combining rate. Note that ϵ_2 is identical to ϵ_1 when $\tau_1 = \tau_2$. The double-exponential filter is rather more well-fitted than the single one for WFBM and DBM (likelihood ratio test, $p \ll 0.001$; adjusted r^2 for double and single exponential filters are 0.99 and 0.98 for WFBM, and 0.94 and 0.85 for DBM). The kernel for WFBM has a negative value around L but it disappears if L is much longer than K. The kernel for DBM drops sharply and decays slowly. The sharp drop probably arose from the exponential decay of reward history, which is embedded in the posterior distributions [Equation (10)]. Because a decision is made due to the difference in two predictive distributions and both distributions decay at the same rate, the effect of one predictive distribution would have persisted slightly longer and hence the kernel included a longer exponential component. This characteristic is qualitatively consistent with the experimental results Corrado et al. (2005). The fitting parameters for the two monkeys in Corrado et al. (2005) were $\rho = 0.4$, $\tau_1 = 2.2$, and $\tau_2 = 17.0$ (monkey F), and $\rho = 0.25$, $\tau_1 = 0.9$, and $\tau_2 = 12.6$ (monkey G). Although there were no suitable WFBM and DBM parameters that exactly matched their fitting parameters to those of the monkeys, similar values were obtained for smaller L and larger α (**Figure 4B**).

It is known that the probability of switching alternatives is nearly constant against the number of consecutive choices for one

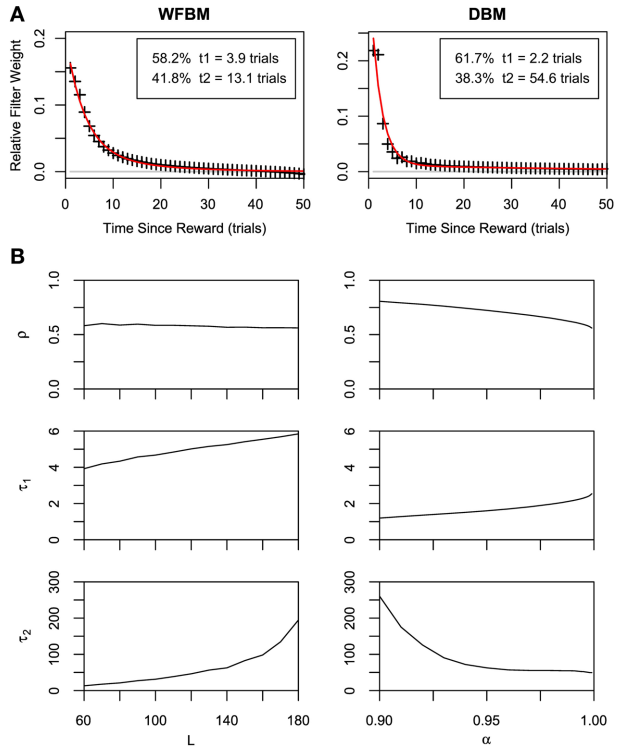

FIGURE 4 | Results of Wiener-Hopf analysis for WFBM and DBM in dynamic foraging task. (A) Symbols represent normalized Wiener-Hopf kernel and red line represents best fitted double-exponential filter. Double-exponential filters are better fitted to data than single-exponential filter (likelihood ratio test, $p \ll 0.001$). Insets show time constants for each exponential component and their combining rate. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. **(B)** Fitted parameters of double-exponential filter ρ, τ_1, and τ_2 to simulation data of WFBM (left column) and DBM (right column). Abscissas represent parameters of WFBM or DBM.

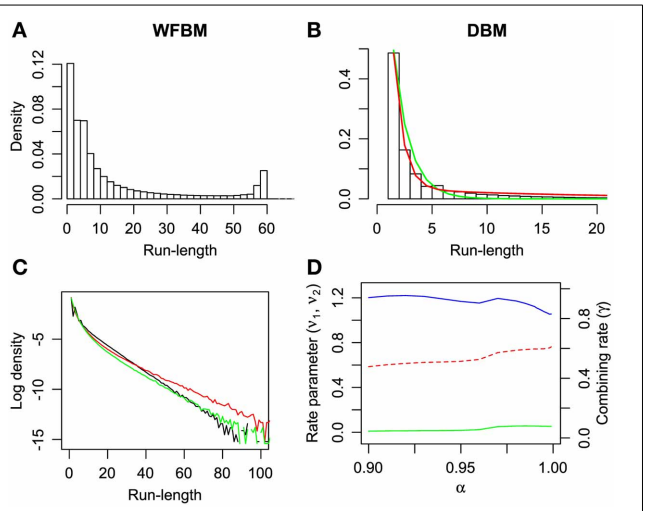

FIGURE 5 | Run-length distributions of windowed FBM and DBM in dynamic foraging task. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. **(A,B)** Bars represent densities of run length for alternative #1. Single (green line) and double-exponential (red line) functions fitted to run-length distributions of DBM. Double-exponential function is fitted better than single one (likelihood ratio test, $p \ll 0.001$). **(C)** Log probability density of run-length distribution of DBM (black line) and linear-nonlinear Poisson models (red and green lines) that are fitted to monkeys' experimental data in Corrado et al. (2005). **(D)** Fitted parameters of double-exponential function with different values of α. Left ordinate indicates value of rate parameters ν_1 (green line) and ν_2 (blue line), and right indicates value of combining rate γ (red line).

alternative (run length) in the concurrent VI schedule (Heyman and Luce, 1979). Hence, run lengths are distributed exponentially but, in a dynamic foraging task, the distribution seems to be a mixture of exponentials (Corrado et al., 2005). The distribution of WFBM does not monotonically decrease and there is a peak where the run length is nearly equal to L. Therefore, the distribution is neither an exponential nor a mixture of exponentials. This nature is consistent on different values of L. However, DBM demonstrates an exponential like distribution. We fitted single and double exponential functions,

$$\begin{aligned}\phi_1(l) &= \nu_0 \exp(-\nu_0(l-1)),\\ \phi_2(l) &= \gamma\nu_1 \exp(-\nu_1(l-1))\\ &\quad + (1-\gamma)\nu_2 \exp(-\nu_2(l-1)),\end{aligned} \tag{13}$$

to the distribution, where $l \geq 1$ is the run length, ν_0 and $\nu_1 < \nu_2$ are the rate parameters and γ is the combining rate. The distribution is well-fitted by the double exponential function (**Figure 5B**; likelihood ratio test, $p \ll 0.001$; r^2 for the double and single exponential functions are 0.99 for the former and 0.96 for the latter). The run-length distribution in monkey experiments has few frequencies of a very short run length; however our models have the largest frequency at the run length of 1 (**Figures 5A,B**). This difference can be due to the absence of change-over-delay (COD) in our schedule. If our model had and exploited prior knowledge about COD as well as the proposed model for the previous experiment (Corrado et al., 2005), the frequency at a run length of 1 could disappear. We simulated linear-nonlinear-Poisson (LNP) models that were fitted to the monkeys' experimental data in Corrado et al. (2005) and compared run-length distributions (**Figure 5C**). Note that COD was not considered for the LNP models that was different from Corrado et al.'s approach Corrado et al. (2005). Because the absence of COD could affect the occurrence of short run lengths, log probability densities were compared to count differences at long run lengths. The calculated mean squared differences of DBM against LNP models for two monkeys corresponded to ∼0.67 and 0.16. The double-exponential function is better than the single one in different α and the fitted parameters are slightly affected by α (**Figure 5D**).

2.2.1. Harvesting performance

Figure 6A compares the harvesting performance of the models, which is normalized by the performance of a near-optimal probabilistic decision making model. The near-optimal model knows the details of the schedules, i.e., both the baiting probabilities and the change points. It distributes its choices according to the choice probabilities that on average maximize the total reward (Sakai and Fukai, 2008a). Due to such given knowledge, none of

the other models can exceed the performance of the near-optimal model. We carried out paired t-tests between the models, in which the means of total reward for an identical schedule were paired. The FBM and WFBM ($L = 60$) are more inferior than the random choice model that chooses by tossing an unbiased coin. The DBM ($\alpha = 0.99$) outperforms FBM, WFBM, and LNP models ($p \ll 0.001$) but the differences from the LNP models are very small. Harvesting performance is less when a model memorizes a more distant past (**Figure 6B**).

3. DISCUSSION

We demonstrated that deterministic Bayesian decision making models can account for the matching law. We confirmed that a simple Bernoulli estimator with a deterministic decision policy demonstrated matching behavior in a static foraging task. We also studied an extended model that includes a belief about a changing environment. The belief effectively works to wipe out the past experience of the model and hence the model can capture three characteristics of behaviors observed in the experiments. First, our model accounts for undermatching, which is a well-known phenomenon in which choices deviate slightly from the matching law (Baum, 1974, 1979; Sugrue et al., 2004). Several studies have addressed possible causes of undermatching, i.e., limitations in the learning rule (Soltani and Wang, 2006), mistuning of parameters (Loewenstein, 2008), and diffusion of synaptic weights (Katahira et al., 2012). This study suggested the cause from a computational perspective, i.e., undermatching was the consequence of a belief in environmental volatility. Second, our model exhibits

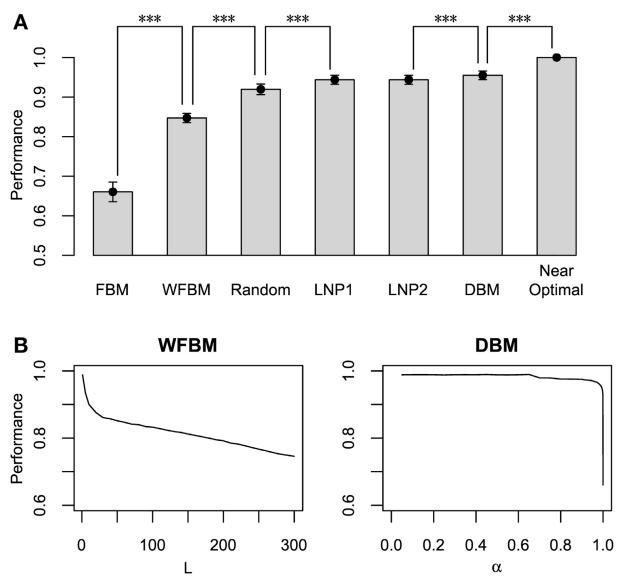

FIGURE 6 | Normalized harvesting performance of each model. (A) Average normalized total rewards earned by each model divided by average total rewards of near-optimal model. Near-optimal model uses strategy that maximizes average total rewards proposed by Sakai and Fukai (2008a) with previous knowledge on details of schedule. Error bars indicate standard deviations around mean. Simulation parameters were set to $L = 60$ and $\alpha = 0.99$. **(B)** Harvesting performance of WFBM and DBM as a function of their parameters. $^{***}p \ll 0.001$.

double-exponential shaped reward history dependency. This is consistent with recent monkey experiments (Corrado et al., 2005; Lau and Glimcher, 2005). Third, the run-length distribution of our model is better fitted by a double-exponential function than a single exponential function. This is also consistent with the previous study (Corrado et al., 2005) although our task did not include changeover delay, which can strongly affect the frequency of shorter run lengths. Quantitatively validating our model such as checking its goodness of fit to raw experimental data would be worthwhile.

The previous models implicitly or explicitly use the strategy of probabilistic choice selection and they learn the choice probability of respective alternatives that satisfy the matching law (Corrado et al., 2005; Lau and Glimcher, 2005; Loewenstein and Seung, 2006; Soltani and Wang, 2006; Sakai and Fukai, 2008a; Simen and Cohen, 2009). Such probabilistic models use a scaling parameter that maps internal decision variables to appropriate choice probabilities and the parameter generally requires fine-tuning (Soltani and Wang, 2006; Fusi et al., 2007). In contrast, as our models act deterministically according to decision variables, no tuning is required for a parameter at the decision stage.

We argued that matching behavior can be explained by a deterministic choice strategy at the computational level. Loewenstein and Seung (2006) proposed biologically inspired synaptic learning rules for neural networks at the neural implementation level. They proved that neural networks developed by covariance-based learning with the assumption of a low learning rate demonstrated matching behaviors. However, this assumption causes the choice to be affected by relatively distant past rewards and the kernel for reward history dependency consequently flattens. A more microscopic spiking neural network model, in which double-exponential dependency in foraging tasks is demonstrated, has been proposed (Soltani and Wang, 2006). However, there is a huge gap between the computational principles of our deterministic macroscopic models and their stochastic microscopic model. This gap can be filled by using a method of reducing spiking neuron models to the diffusion equation (Roxin and Ledberg, 2008). There have been some other neural network models that can show heavy-tailed dependency of choices on past experience. A reservoir network (Jaeger et al., 2007), which can reproduce neural activity in the monkey prefrontal cortex, preserves the memory trace of a reward with one or two time constants (Bernacchia et al., 2011). The composite learning system of faster and slower components is flexible to abrupt changes in the environment (Fusi et al., 2007). These models could be a possible neural implementation for our model. Furthermore, our models are an extension of that by Yu & Cohen who argued that decision variables of their model can be approximated by a linear exponential filter, and that there are neural implementations for that operation (Yu and Cohen, 2009).

Because matching behavior often deviates from optimal behavior in the sense of total reward maximization (Vaughan, 1981), it is not likely to be a consequence of optimization. However, our model acts optimally in terms of Bayesian decision making with an incorrect assumption about the environment, indicating that matching behavior is a bounded optimal behavior. This idea is consistent with the theory of Sakai and Fukai (2008b)

who found any learning method neglecting the effect of a choice on future rewards displays matching behavior if choice probabilities are differentiable with respect to parameters (Sakai and Fukai, 2008b). Note that the choice probabilities of our model are not differentiable. Hence, we confirmed that their theory could be correct in such extreme cases.

4. MATERIALS AND METHODS

4.1. DETAILS OF SIMULATION

The reward schedule is analogous to the experiment by Corrado et al. (2005). We randomly set the baiting probabilities that satisfied $\lambda_1 + \lambda_2 = 0.3$ and their ratios were 1:8, 1:6, 1:3, 1:2, 1:1, 2:1, 3:1, 6:1, and 8:1 in a static setting. There were 10, 000 trials in the simulations. The baiting schedule in the dynamic setting was divided into blocks, in which the baiting probabilities were fixed, and their sum and ratios were the same as those in the static setting. The block length was uniformly sampled from [50, 300] and there were 300 blocks in the simulations. We did not include change-over-delay (COD), i.e., the cost to switch from one alternative to another, which was different from Corrado et al. (2005). The hyper-parameters were set to $a = 1$ and $b = 1$ in all the simulations.

ACKNOWLEDGMENTS

This work was partially supported by a Grant-in-Aid from the Japan Society for the Promotion of Science (JSPS) Fellows of the Ministry of Education, Culture, Sports, Science and Technology (No. 11J06433).

REFERENCES

Anderson, K. G., Velkey, A. J., and Woolverton, W. L. (2002). The generalized matching law as a predictor of choice between cocaine and food in rhesus monkeys. *Psychopharmacology* 163, 319–326. doi: 10.1007/s00213-002-1012-7

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *J. Exp. Anal. Behav.* 22, 231–242. doi: 10.1901/jeab.1974.22-231

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *J. Exp. Anal. Behav.* 32, 269–281. doi: 10.1901/jeab.1979.32-269

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *J. Exp. Anal. Behav.* 36, 387–403. doi: 10.1901/jeab.1981.36-387

Baum, W. M. (1982). Choice, changeover, and travel. *J. Exp. Anal. Behav.* 38, 35–49. doi: 10.1901/jeab.1982.38-35

Baum, W. M., and Rachlin, H. C. (1969). Choice as time allocation1. *J. Exp. Anal. Behav.* 12, 861–874. doi: 10.1901/jeab.1969.12-861

Baum, W. M., Schwendiman, J. W., and Bell, K. E. (1999). Choice, contingency discrimination, and foraging theory. *J. Exp. Anal. Behav.* 71, 355–373. doi: 10.1901/jeab.1999.71-355

Bernacchia, A., Seo, H., Lee, D., and Wang, X. J. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.* 14, 366–372. doi: 10.1038/nn.2752

Corrado, G. S., Sugrue, L. P., Seung, H. S., and Newsome, W. T. (2005). Linear-nonlinear-poisson models of primate choice dynamics. *J. Exp. Anal. Behav.* 84, 581–617. doi: 10.1901/jeab.2005.23-05

Davison, M., and McCarthy, D. (1988). *The Matching Law: A Research Review*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

de Villiers, P. A., and Herrnstein, R. J. (1976). Toward a law of response strength. *Psychol. Bull.* 83, 1131–1153. doi: 10.1037/0033-2909.83.6.1131

Fusi, S., Asaad, W. F., Miller, E. K., and Wang, X. J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron* 54, 319–333. doi: 10.1016/j.neuron.2007.03.017

Gallistel, C. R. (1994). Foraging for brain stimulation: toward a neurobiology of computation. *Cognition* 50, 151–170. doi: 10.1016/0010-0277(94)90026-4

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *J. Exp. Anal. Behav.* 4, 267–272. doi: 10.1901/jeab.1961.4-267

Heyman, G. M., and Luce, R. D. (1979). Operant matching is not a logical consequence of maximizing reinforcement rate. *Learn. Behav.* 7, 133–140. doi: 10.3758/BF03209261

Hinson, J. M., and Staddon, J. E. R. (1983). Matching, maximizing, and hill-climbing. *J. Exp. Anal. Behav.* 40, 321–331. doi: 10.1901/jeab.1983.40-321

Jaeger, H., Lukoeviius, M., Popovici, D., and Siewert, U. (2007). Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.* 20, 335–352. doi: 10.1016/j.neunet.2007.04.016

Katahira, K., Okanoya, K., and Okada, M. (2012). Statistical mechanics of reward-modulated learning in decision-making networks. *Neural Comput.* 24, 1230–1270. doi: 10.1162/NECO_a_00264

Lau, B., and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* 84, 555–579. doi: 10.1901/jeab.2005.110-04

Loewenstein, Y. (2008). Robustness of learning that is based on covariance-driven synaptic plasticity. *PLoS Comput. Biol.* 4:e1000007. doi: 10.1371/journal.pcbi.1000007

Loewenstein, Y., and Seung, H. S. (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15224–15229. doi: 10.1073/pnas.0505220103

Roxin, A., and Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Comput. Biol.* 4:e1000046. doi: 10.1371/journal.pcbi.1000046

Sakai, Y., and Fukai, T. (2008a). The actor-critic learning is behind the matching law: matching versus optimal behaviors. *Neural Comput.* 20, 227–251. doi: 10.1162/neco.2008.20.1.227

Sakai, Y., and Fukai, T. (2008b). When does reward maximization lead to matching law? *PLoS ONE* 3:e3795. doi: 10.1371/journal.pone.0003795

Simen, P., and Cohen, J. D. (2009). Explicit melioration by a neural diffusion model. *Brain Res.* 1299, 95–117. doi: 10.1016/j.brainres.2009.07.017

Soltani, A., and Wang, X. J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J. Neurosci.* 26, 3731–3744. doi: 10.1523/JNEUROSCI.5159-05.2006

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science* 304, 1782–1787. doi: 10.1126/science.1094765

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat. Rev. Neurosci.* 6, 363–375. doi: 10.1038/nrn1666

Vaughan, Jr. W. (1981). Melioration, matching, and maximization. *J. Exp. Anal. Behav.* 36, 141–149. doi: 10.1901/jeab.1981.36-141

Yu, A. J., and Cohen, J. D. (2009). "Sequential effects: superstition or rational behavior," in *Advances in Neural Information Processing Systems* 21, 1873–1880. Available online at: http://books.nips.cc/nips21.html

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2013; accepted: 05 February 2014; published online: 04 March 2014. Citation: Saito H, Katahira K, Okanoya K and Okada M (2014) Bayesian deterministic decision making: a normative account of the operant matching law and heavy-tailed reward history dependency of choices. Front. Comput. Neurosci. ***8****:18. doi: 10.3389/fncom.2014.00018*

This article was submitted to the journal Frontiers in Computational Neuroscience.