Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Modeling and analysis of dynamic decision making in sequential two-choice tasks

Linh Vu Kristi A. Morgansen

***Abstract*—We present the construction and analysis of a dynamical system model for human decision making in sequential two-choice tasks, in which a human subject makes a series of interrelated decisions between two choices in order to obtain the maximum reward. For a nominal decision making policy inspired by behavioral aspects of humans, we show asymptotic behavior of such decision making process in sequential two-choice tasks for various types of reward structures. Our work gives a control theory oriented perspective to the experiments carried out by cognitive scientists.**

I. Introduction

Operations of mixed teams of humans and robots have recently attracted new research interests with the goal of ensuring that the entire human/robot system maintains certain required performance despite that performance of humans is affected by physiological and psychological factors and skill levels (see, e.g., [1]). One way to address the above issue is to incorporate human decision making dynamics into the design of autonomous robots. This line of research requires knowledge of both human decision making and autonomous control and has attracted collaborative research of cognitive researchers and control researchers.

In this paper, we use control theoretic tools to study a particular type of human decision making, namely decision making in *sequential two-choice tasks* [2], [3]. This type of task is specifically devised by cognitive scientists to study cognitive and behavioral aspects of human decision making [2], [3] (which were inspired by foraging behaviors of honeybees [4]).

Human decision making in sequential two-choice tasks has been modeled using the prediction-error model [3], which is a neural network/reinforcement learning model with two weights, one for each choice, where the probability of the next choice is biased as a function of the difference between the weights. Also closely related to sequential two-choice tasks are two alternative forced-choice tasks, such as the task of determining the direction of moving dots embedded in a noisy screen. Two alternative forced-choice tasks have been modeled using drift diffusion models (see, e.g. [5] and the references therein), in which the difference between the amounts of evidence supporting one choice over the other is integrated over time until it crosses a threshold.

This work aims to bring a control systems perspective to the modeling and analysis of sequential two-choice tasks. In particular, we focus on *asymptotic properties* of human decision making, a feature which has not been addressed adequately in [3], [5], [6]. While human decision making is stochastic in nature, as a first step, we model human decision making as a deterministic policy motivated by two basic human behavioral characteristics. We then discuss extensions of such nominal policy to capture other human behavioral aspects. Compared to the prediction-error model, our model is somewhat simpler but it enables us to analyze asymptotic behaviors for various types of reward structures in sequential two-choice tasks (see also [7] for another approach to the modeling and analysis of sequential two-choice tasks as well as how to map this type of task to certain operations of mixed teams of humans and robots).

Our work here lies in between the areas of cognitive science and control systems theory. With respect to the cognitive research community, this work brings another perspective on and explanation of the dynamics of human decision making. To the control community, this work introduces a particular class of dynamical systems coming from cognitive science in which the dynamics to be controlled are completely unknown to the controller, the controller has limited memory and computational capability, and the system's behavior is then assessed via structure of output functions.

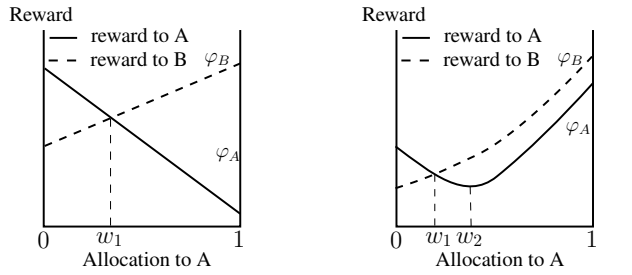
II. Sequential two-choice task

In a sequential two-choice task [2], a participant is presented with two choices (e.g., two buttons), A and B. At each time, the participant can choose either A or B and receives a reward afterward. The goal is

This work is supported in part by AFOSR grant FA9550-07-10528.

Linh Vu and Kristi A. Morgansen are with the department of Aeronautics and Astronautics, University of Washington, Seattle. Email: {linhvu,morgansn}@u.washington.edu.

to maximize the reward. The reward is calculated as follows: if the participant chooses A, and the percentage of A among the last fixed number of (e.g. 20) decisions is x, the reward is $\varphi_A(x)$; if the participant chooses B, and the percentage of A among the last 20 decisions is x, the reward is $\varphi_B(x)$ (see Fig. 1). The number 20 is defined as the window of the task. The pair (φ_A, φ_B) is called a *reward structure*. Examples of common reward structures are plotted in Fig. 1, where the horizonal axis is the percentage of allocation to A in last 20 decisions, and the vertical axis is the reward (which is a non-negative number) for choice of A or B at a given percentage of A.



a. Matching shoulder task b. Rising optimum task

Fig. 1. Common reward structures [3].

The difficulty for a participant in the task is that the participant knows neither how the reward is calculated nor the reward functions φ_A and φ_B. Cognitive scientists are interested in finding out whether a human can find the global optimum for a given reward structure (e.g., in Fig 1b, the global optimum decision is near the point 1 of the horizontal axis) and how human characteristics (such as personality) influence performance in these tasks.

In experiments using the matching shoulder task (Fig. 1a) [2], cognitive scientists observed that for a majority of the human test subjects (18 out of 24), the human decisions were such that the average of A over the entire experiment was biased towards the point w_1 of the horizontal axis (which is known as a *crossing point* in the cognitive literature [8]). In the rising optimum task (Fig. 1b), for about half of all participants (14 out of 25), the average of A over the entire experiment was biased toward the point w_1. A smaller percentage of the participants were biased toward the point 1 of the horizontal axis. These particular biases suggest that there is some mechanism underlying human decision making in sequential two-choice tasks.

III. Modeling decision making in sequential two-choice tasks

Denote by $t_k \in \mathbb{R}, k = 0, 1, \ldots$ the times at which the human makes decisions (where t_0 is the starting time), and by $u(t_k)$ the corresponding decisions, where $u(t_k) \in \{A, B\}$. Denote by x the fraction of A in the last N choices, where N is the window of the task. Between t_k and t_{k+1}, x does not change, and so the dynamics of x are $\dot{x}(t) = 0$ for $t \in (t_k, t_k + 1)$. At time t_k, x is changed[1] according to the following *impulsive map* g (also known as a reset map):

$$\begin{aligned} x(t_k) &= g(x(t_k^-), u(t_k)) \\ &=: \begin{cases} x(t_k^-) + \frac{1}{N} & \text{if } u(t_k) = A, \\ & \quad u(t_k) \neq u(t_{k-N}) \\ x(t_k^-) - \frac{1}{N} & \text{if } u(t_k) = B, \\ & \quad u(t_k) \neq u(t_{k-N}) \\ x(t_k^-) & \text{if } u(t_k) = u(t_{k-N}). \end{cases} \end{aligned} \tag{1}$$

Let (φ_A, φ_B) be a reward structure, where $\{\varphi_A, \varphi_B\} : [0, 1] \to [0, R_{max}]$ for some number R_{max}. The reward to the participant at time t_k is

$$y(t_k) = \varphi_{u(t_k)}(x(t_k)). \tag{2}$$

In control terminology, the decision u is the control signal, and y is the output. In general, the control signal is of the form

$$u(t_k) = \rho(I(t_k)) \tag{3}$$

for some function ρ (which could be probabilistic and vary from person to person), and $I(t_k) := \{u(t_i), y(t_i), 0 \leqslant i < k\}$ is the information available to the controller (i.e. the human in this case) at time t_k. The closed loop system of the sequential two choice task is written as

$$\begin{cases} \dot{x}(t) &= 0 \quad t \neq t_k \\ x(t_k) &= g(x(t_k^-), u(t_k)) \\ y(t_k) &= \varphi_{u(t_k)}(x(t_k)) \\ u(t_k) &= \rho(I(t_k)) \\ I(t_k) &= I(t_{k-1}) \cup \{y(t_{k-1}), u(t_{k-1})\} \end{cases} \tag{4}$$

with some initial values for $u(t_{-N}), \ldots, u(t_{-1})$, the initial state $x(t_{-1})$ being the number of A in $\{u(t_{-N}), \ldots, u(t_{-1})\}$, and $I(t_{-1}) = \emptyset$, where g is as in (1), φ is as in (2), and ρ is as in (3).

[1]Without loss of generality, we assume that x is continuous from the right.

We chose to treat x as a continuous-time variable because in general, the variable x, which represents the state of the environment with which the human interacts, can have its own dynamics, i.e. $\dot{x} = 0$ in (4) can be generalized to $\dot{x}(t) = f(x(t))$ (though in this paper, we only consider the sequential two-choices tasks represented by (4)). Fig. 2 depicts the closed loop of sequential two-choice tasks in general.

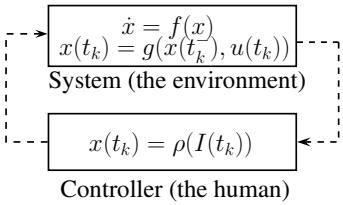


Fig. 2. The closed loop system of a sequential decision task. The dash line is to indicate that communication happens only at times t_k.

A. The γ-policy for decision making

We do not solve for a control law ρ which optimally plays the sequential two-choice tasks, but rather, try to understand human decision making in such tasks (which could be non-optimal). We seek a static control law ρ, which is termed a *policy* in this context (in cognitive science, such approach is called a top-down approach, and policies are called normative strategies; see, e.g., [2]). A static ρ is of interest for the following reason: a static ρ can be seen as an input-output characterization of human decision making dynamics. Such a static ρ makes analysis of asymptotic behaviors of the closed loop system (4) possible. Also, this input-output approach allows us to incorporate human behavioral characteristics (such as exploring or hedging) directly into the model (see Section V).

We now present one simple decision making rule—termed the γ-policy—in sequential two-choice tasks:

$$u(t_k) := \begin{cases} u(t_{k-1}) & \text{if } y(t_{k-1}) \geqslant y(t_{k-2}) \\ \text{switch}(u(t_{k-1})) & \text{else,} \end{cases} \tag{5}$$

where $\text{switch}(a) = B$ if $a = A$, and $\text{switch}(a) = A$ if $a = B$. By convention, $y(t_{-1}) = \varphi_{u(t_{-1})}(x(t_{-1}))$.

The rationale behind the γ-policy is as follows: In sequential two-choice tasks, we postulate that humans largely follow the following two courses of action:

- A1: If the last reward increases or does not decrease, one keeps the current choice.
- A2: If the last reward decreases, one will immediately switch the decision.

The above courses of action are illuminated by human behavioral characteristics. If the outcome is as expected i.e. when the last choice increases the reward, there is no incentive to change from the last decision (alternatively, it is the incentive to continue with the same decision). If the outcome is not as expected i.e when the last choice decreases the reward, it is the incentive to switch the decision (in order to avoid further potential losses)[2].

In cognitive psychology terminology, actions of Type A1 are known as *exploitation*, and actions of Type A2 are known as *exploration* (in fact, relationships between exploitation and exploration in human decision making is a major research theme in cognitive science [11]). Using this terminology, the first case of (5) encodes exploitation, and the second case encodes exploration. Switching between the two is triggered by whether expectation is met or not.

The actions A1 and A2 are idealistic. Some humans may try to explore by switching decisions even if there is no decrease in rewards (for example, in gambling type of activities), or some may stay with a decision even if there is a temporary decline in rewards (for example, those traits can be found in long term investors). We will discuss modifications of the policy (5) to incorporate these deviations later in Section V, but for the moment, for the sake of conveying our idea and for analysis, we use the (ideal) policy (5).

IV. Analysis

A. Reward structures

Reward structures are continuous functions as constructed in the experiments carried out by cognitive scientists. However, the outcomes of the experiments depend only on values of the reward functions at the discrete positions $X = \{0, 1/N, \ldots, N-1/N, 1\}$.

In this paper, we consider eight basic types of reward structures based on the monotonicity of φ_A and φ_B and their relative positions. We write $\varphi_A > \varphi_B$ if $\varphi_A(x) > \varphi_B(x)\ \forall x \in X$ (and similarly for $\varphi_A < \varphi_B$). The definitions and examples the basic reward structures are plotted in Fig. 3. The range of the rewards is $[0, R_{max}]$ for some R_{max}, and without loss of generality, R_{max} can be taken as 1 after some scaling. For a structure $\Gamma = (\varphi_A, \varphi_B)$, denote by $\texttt{Type}(\Gamma)$ the type of Γ.

[2]The γ-policy is very similar in spirit to the win-stay-lose-change strategy in the psychology literature [9] or game theory literature [10] but here the slight difference is that the reward can be the same instead of a clear win-or-lose situation, and there is only one player.

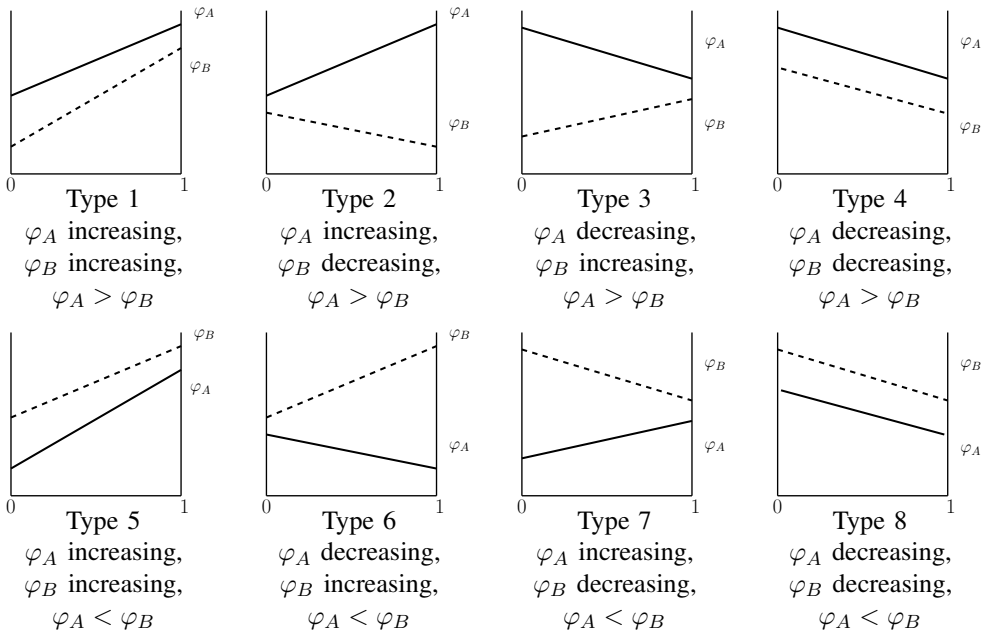


Fig. 3. Basic reward structures for sequential two-choice tasks. The horizontal axis is the percentage of A choices in the last N decisions, and the vertical axis is the rewards for choice of A or B at given percentage of A

To facilitate analysis, we allow the domain of reward structures be general intervals $[a, b] \subset \mathbb{R}$, and the set X is replaced by $[a, a+\Delta, \ldots, b]$ for some $0 < \Delta < b-a$. The definition of the basic reward structures in Fig. 3 unchanges when we replace $[0, 1]$ by an interval $\mathcal{D}$.

B. Asymptotic behavior

We can show that under the γ-policy, the decisions u exhibit an asymptotic behavior. As in the previous section, for generality, we allow the domain of the reward structure be an arbitrary interval $[a, b]$, and further, we allow a general step change Δ instead of $1/N$ in the impulsive map g in (1):

$$x(t_k) = \bar{g}(x(t_k^-), u(t_k)) =: \begin{cases} \min\{x(t_k^-) + \Delta, b\} & \text{if } u(t_k) = A, \\ & u(t_k) \neq u(t_{k-N}) \\ \max\{x(t_k^-) - \Delta, a\} & \text{if } u(t_k) = B, \\ & u(t_k) \neq u(t_{k-N}) \\ x(t_k^-) & \text{if } u(t_k) = u(t_{k-N}). \end{cases} \tag{6}$$

The max and min are to ensure that x is between a and b. Let $z(t_k) := (u(t_{k-1}), u(t_k))$ be the ordered sequence of the two consecutive decisions at time t_{k-1} and t_k; it is clear that $z(t_k) \in \{AA, AB, BA, BB\}$. We write $x \rightarrowtail x^*$ (respectively, $x \rightarrowtail \mathcal{S}$) if there exists $T < \infty$ such that $x(t) = x^*$ (respectively, $x(t) \in \mathcal{S}$) $\forall t \geqslant T$.

Lemma 1 *Consider the system* (4) *with the impulsive map* (6)*. Let* Γ *be a basic reward structure in Fig. 3 on an interval* $[a, b]$*. Under the* γ*-policy* (5)*,*

- *if* $\texttt{Type}(\Gamma) = 1$*, then* $x \rightarrowtail a$ *if* $x(t_{-1}) = a$ *and* $z(t_1) = BB$*, and* $x \rightarrowtail b$ *else*
- *if* $\texttt{Type}(\Gamma) \in \{2, 7\}$*, then either* $x \rightarrowtail a$ *or* $x \rightarrowtail b$
- *if* $\texttt{Type}(\Gamma) \in \{3, 4, 5, 6\}$*, then there exists* $x^* \in X$ *such that* $x \rightarrowtail \{x^*, x^* + \Delta\}$
- *if* $\texttt{Type}(\Gamma) = 8$*, then* $x \rightarrowtail b$ *if* $x(t_{-1}) = b$ *and* $z(t_1) = AA$*, and* $x \rightarrowtail a$ *else.*

Sketch of proof: We construct finite state machines for the variable z as in Fig. 4. Then by examining the graph cycles for each finite state machine, we are able to find patterns for the sequence of decisions. For example, for Type 1 basic reward structures, from its finite state machine, we conclude that unless the original state is starting at $x = a$ and the initial z is BB, all other z will lead to a cycle of decisions AA, and hence, x will approach b in finite time. The analysis is more complicated for other finite state machines which have multiple cycles and cycles with length greater than 1. Due to the space limitation, the proof is omitted; see [12] for details. □.

Corollary 1 *Consider the system* (4) *with the impulsive map* (6)*. Let* Γ *be a basic reward structure on* $[a, b]$*. Under the* γ*-policy* (5)*, for every initial state* $x(0) \in$

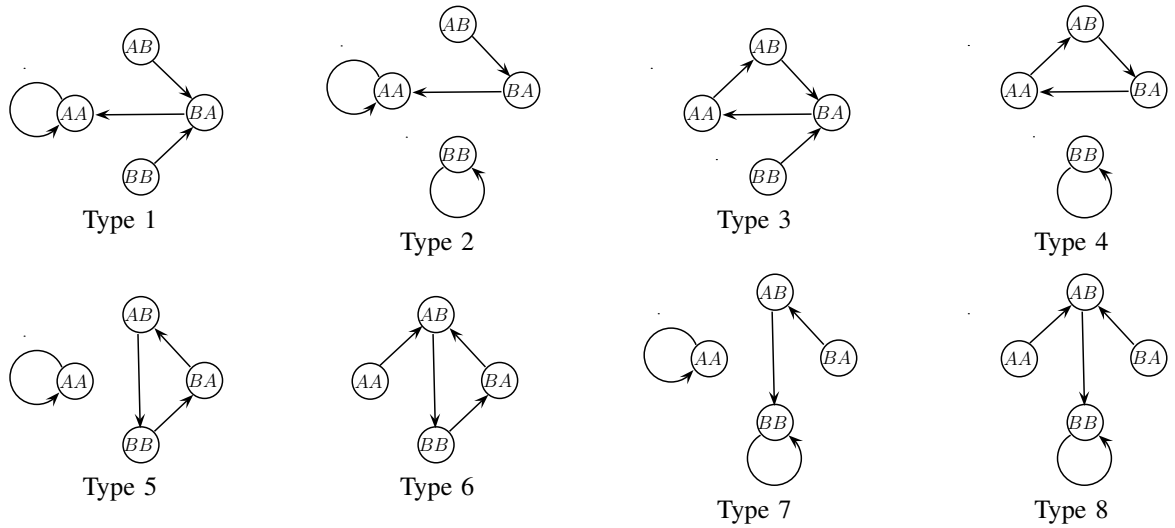


Fig. 4. Finite state machines for the basic reward structures in Fig. 3.

$[a, b]$, *there exists a finite time* T *and* $x^* \in X$ *such that* $x \to \{x^*, x^* + \Delta\}$.

For reward structures not one of the eight basic types in Fig 3, if we can decompose the reward structure into substructures of the basic types in Fig 3, then we can examine the behavior of the overall structure using Lemma 1. For a reward structure Γ on the domain $[0, 1]$, let $\{\Gamma_1, \ldots, \Gamma_m\}$ be the *substructure decomposition* of Γ such that Γ is the concatenation of $\Gamma_1, \ldots, \Gamma_m$. Assume that $\Gamma_1, \ldots, \Gamma_m$ are of the basic types in Fig 3. It can be shown that if such assumption holds, then the decomposition $\Gamma_1, \ldots, \Gamma_m$ is unique.

Example 1 The reward structure in Fig. 1a. can be uniquely decomposed into the substructures Γ_1 on $[0, w_1]$ and Γ_2 on $[w_1, w]$, where w_1 is the intersection of φ_A and φ_B. We have $\texttt{Type}(\Gamma_1) = 3$ and $\texttt{Type}(\Gamma_2) = 6$. $\lhd$

Using Lemma 1, we obtain the following theorem.

Theorem 1 *Consider the system* (4) *under the* γ*-policy* (5) *with a reward structure* Γ *on* $[0, 1]$ *and a window* N. *Suppose that* Γ *can be decomposed into substructures of the eight basic types in Fig. 3. For any initial values of* $u(t_{-N}), \ldots, u(t_{-1})$, *there is a time* $T < \infty$ *and* $x^* \in [a, b]$ *such that* $|x(t) - x^*| \leqslant 1/N$ *for all* $t \geqslant T$.

V. Discussions

The γ–policy is idealistic, while undoubtedly, human behavioral characteristics are rich and diverse. To capture other human factors, the γ–policy can be modified and extended in several ways as described belows.

1) Nondeterministic γ*-policy:* Humans may not follow the γ-policy if one is not consciously aware of it. However, as we discussed in Section III-A, humans may unconsciously follow the γ-policy most of the time due to common human psychology. We can relax the condition that a human follows the γ-policy at all times by allowing the human to only choose the decision dictated by the γ-policy with a probability p. We call $1 - p$ the *error probability*.

In particular, a probabilistic γ-policy is

$$u(t_k) := \begin{cases} u(t_{k-1}) & \text{if } y(t_{k-1}) \geqslant y(t_{k-2}), \\ & \text{with prob. } p \\ \text{switch}(u(t_{k-1})) & \text{if } y(t_{k-1}) < y(t_{k-2}), \\ & \text{with prob. } p \\ \text{switch}(u(t_{k-1})) & \text{if } y(t_{k-1}) \geqslant y(t_{k-2}), \\ & \text{with prob. } 1-p \\ u(t_{k-1}) & \text{if } y(t_{k-1}) < y(t_{k-2}), \\ & \text{with prob. } 1-p. \end{cases} \tag{7}$$

We conjecture that for certain types of basic reward structures, there exists a $p^* \in (0, 1)$ such that for all $p < p^*$, we will recover the deterministic case with probability 1, i.e. for all $p < p^*$, for any initial values $u(t_{-N}), \ldots, u(t_{-1})$, $\exists x^* \in [0, 1]$ such that $P(E|x(t) - x^*| \leqslant 1/N \ \forall t \geqslant T \text{ for some } T) = 1$.

2) Personality: A γ-policy with a threshold δ is:

$$u(t_k) := \begin{cases} u(t_{k-1}) & \text{if } y(t_{k-1}) \geqslant y(t_{k-2}) + \delta \\ switch(u(t_{k-1})) & \text{else.} \end{cases} \tag{8}$$

If $\delta < 0$, the modified strategy (8) means that the behavior is more exploitative: the action does not change unless the reward degrades by an amount of at least δ (this action means a risky behavior because the player does not react to negative development). On the other hand, if $\delta > 0$, the strategy is more explorative: the decision is switched unless the reward is improved by a amount of at least δ (this action can be seen as conservative in sequential two-choice tasks because the goal is to seek optimum, and so it is not risky to explore). Thus, in essence, the variable δ may capture the risk attitude of humans in sequential two-choice tasks: $\delta = 0$ means risk neutral (most people), $\delta > 0$ means conservative behavior (few people), and $\delta < 0$ means risky behavior (few people). We can also have time-varying δ instead of a constant one.

We conjecture that for a certain $\delta > 0$ and certain types of reward structures, the extended γ-policy can mimic the behavior of the explorative type of people who would navigate through a temporary dip to reach the optimum (this behavior has also been observed in experiments).

3) Response time: The effect of response time under pressure (e.g. deadline) can be included as degradation of the reward after a certain time window (alternatively, one can cast this aspect as a changing environment). A participant could be told that one has τ unit of times to make a decision after which the reward will deteriorate. This type of tasks can be captured, for example, by the following output function

$$y(t_k) = \begin{cases} \varphi_{u(t_k)}(x(t_k)) & \text{if } t_k - t_{k-1} < \tau \\ \varphi_{u(t_k)}(x(t_k)) e^{-\lambda(t_k - t_{k-1})} & \text{if } t_k - t_{k-1} > \tau \end{cases} \tag{9}$$

for some $\lambda > 0$.

The effect of response time under deadline pressure can also be studied under the probabilistic γ-policy framework, in which the error probability $p_e = 1 - p$ is a function of τ. Reasonable relationships between p_e and τ are such that p_e is larger if τ is smaller, p_e is smaller if τ is larger, and probably, $p_e \to 0$ as $\tau \to \infty$. Different types of relationship between p_e and τ could exist depending human personality. This modification also captures the issue between decision time and error rate (speed vs. accuracy) in human decision making.

4) Longer memory: Another aspect is to include more memory in the γ-policy (for example, chess players can remember a larger number of steps than average people). Instead of using the last two rewards in the γ-policy, one can have a policy with three or more past rewards.

VI. Conclusions

We modeled human decision making in sequential two-choice tasks as closed-loop dynamical systems. Using a decision making policy known as the γ-policy, we showed that under such policy, the percentage of A asymptotically converges to a point or a ball of radius $1/N$, where N is the window size. Future work is to cover basic reward structures with flat reward functions, to explore the extension discussed in Section V, and to validate the theoretical results with data from experiments with human subjects.

References

[1] Board on Mathematical Sciences and Their Applications, *Basic Research in Information Science and Technology for Air Force Needs*. The National Academies Press, 2006.

[2] D. Egelman, C. Person, and P. Montague, "A computational role for dopamine delivery in human decision-making," *J. Cogn. Neurosci*, vol. 10, pp. 623–630, 1998.

[3] P. R. Montague and G. S. Berns, "Neural economics and the biological substrates of valuation," *Neuron*, vol. 36, pp. 265–284, 2002.

[4] L. A. Real, "Animal choice behavior and the evolution of cognitive architecture," *Science*, vol. 253, pp. 980–986, 1991.

[5] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks," *Psychological Review*, vol. 113, no. 4, pp. 700–765, 2006.

[6] R. Bogacz, S. M. McClure, J. Li, J. D. Cohen, and P. R. Montague, "Short-term memory traces for action bias in human reinforcement learning," *Brain Research*, vol. 1153, pp. 111–121, 2007.

[7] M. Cao, A. R. Stewart, and N. E. Leonard, "Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis," in *Proceedings of the 47th IEEE Conf. Decision and Control*, 2008.

[8] R. J. Herrnstein, "Rational choice theory: necessary but not sufficient," *American Psychologist*, vol. 45, pp. 356–367, 1990.

[9] H. H. Kelley, J. W. Thibaut, R. Radloff, and D. Mundy, "The development of cooperation in the minimal social situation'," *Psychological Monographs*, vol. 76, no. 19, 1962.

[10] R. H., "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, pp. 527–535, 1952.

[11] J. D. Cohen, S. M. McClure, and A. J. Yu, "Should I stay or should I go? How the human brain manages the tradeoff between exploitation and exploration." *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, pp. 933–942, 2007.

[12] L. Vu and K. A. Morgansen, "Modeling and analysis of dynamic decision making in sequential two-choice decision tasks," *Preprint. http://vger.aa.washington.edu/~linhvu/research/two_choice_tasks.pdf*, 2008.