Statistical Properties of Decision-Making Governed by Reinforcement Learning with Status Quo Bias

Ihor Lubashhevsky and Kosuke Hijikata
University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu City, Fukushima 965-8560, Japan
E-mails for contacts: (IL) i-lubash@u-aizu.ac.jp

Abstract

Within the paradigm of human intermittent control over unstable systems human behavior admits the interpretation as a sequence of point-like moments when the operator makes decision on activating or deactivating the control. These decision-making events are assumed to be governed by the information about the state of system under control which the operator accumulates continuously. In the present work we propose the concept of reinforcement learning with decision inertia (the status quo bias) that opens a gate to applying the formalism of reinforcement learning to describing human intermittent control. The basic feature of such reinforcement learning is that human behavior in a sequence of selecting available options exhibits quasi-continuous dynamics. Numerical simulation based on a fairly simple model demonstrates that the proposed formalism does possess the required properties of quasi-continuous behavior.

1 Human Intermittent Control as a Stream of Decision-Making Events

During the last decades there has been developed a novel paradigm of describing human actions in controlling various unstable systems—event driven human intermittent control (for a review see, e.g., [1, 2, 3, 4, 5]). The balancing of inverted pendulum or human body upright position exemplifies it. The characteristic feature of this control is that human actions in keeping a controlled system near a desired position form an alternate sequence of active and passive phases instead of being continuous through the process. During the passive phase the actions of human operator in governing the system motion are halted. Broadly speaking, the operator accumulates the information about the system motion without affecting it or changing the current state of control. The individual fragments of active phase admit interpretation in terms of the open-loop control (see, e.g., [2]). In other words, having started the system correction, the operator practically does not respond to local changes in the system motion until the planned action is not implemented and the system is returned to a certain neighborhood of its desired position.

According to the modern state of the art in the theory of human intermittent control (see, e.g., [4, 6, 7, 8]) these human actions are assumed to be rather deterministic. It implies that the control is active when the discrepancy between the desired and actual system states exceeds a certain threshold. Naturally some noise can be added to smooth this step-wise transition in human actions.

Recently [9, 10] based on data collected in experiments on the balancing of inverted pendulum with overdamped dynamics we proposed a new concept of intermittent control activation—the noise-driven activation—which considers human actions highly irregular. In this case human response to the discrepancy between the desired and actual system states should be treated as substantially probabilistic events. The gist of the proposed approach to modeling human intermittent control is the interpretation of human actions as a sequence of point-like events when the operator decides to activate the control and, then, to halt its implementation. Between these events the operator behavior may be regarded as passive in the sense that he either does not correct the current system motion (passive phase) or does not change the already initiated actions of system correction (passive phase). Only the transitions from passive to active phases and, vice verse, from active to passive phases are treated as the intentional actions of the operator caused by his response to changes in the system motion. In other words, these point-like events are conceived of as the instants when the operator makes decision to change the current state of his control (i.e., to activate or halt the control action depending on the information about the system dynamics accumulated between the decision-making events). This accumulation of information about the system state needs a certain finite time because of the bounded capacity of human cognition and the delay in information processing.

Reinforcement Learning and Decision Inertial

In our on-going research we intent to develop a theory of human intermittent control turning to the noted concept of point-like events of the operator decision-making. The reinforcement learning paradigm seems to provide us with an appropriate formalism. Namely, reinforcement learning may be treated as the process of exploring unknown, changeable environment with a feedback mechanism (see, e.g., [11]), which is illustrated
in Fig. 1. In a simplified form it is the algorithm comprising the following three elements:

E1: An agent interacts with an environment and via the trial-error strategy accumulates the information about its states from the standpoint of their relevance to a certain goal pursued by the given agent. The preference of each state $i$ is quantified by the value $q_i$ being the result of many trials of choosing a given state $i$. Each time $t$ a state $i$ is chosen (currently explored) the agent receives a reward $r_i(t)$. It is described by the following equation

$$q_i(t + \tau) = (1 - \epsilon)q_i(t) + r_i(t),$$

where the coefficient $0 \leq \epsilon \leq 1$ describes the rate at which the information decays as time goes on and $\tau$ is the time span between successive events in exploring the environment. If currently a given state $i$ is not chosen for exploring its priority value undergoes only the time decay described as:

$$q_i(t + \tau) = (1 - \epsilon)q_i(t).$$

It is a rather simple model for reinforcement learning where many effects are not taken into account. For example, in estimating the preference of currently non-chosen states the information from the previous time moments can be used. Nevertheless, for our purpose this simplest model can be accepted without lost of generality.

E2: The higher the priority value, the more probable the agent choice of the corresponding state. Accepting that the priority value can be introduced only for situations when the choice between possible options is available, the agent choice of a state $i$ is describe by the probability $p_i$

$$p_i(t) = \frac{\exp\{q_i(t)\}}{\sum_{k=1}^{N} \exp\{q_k(t)\}}.$$  

Here the sum $\sum_{k=1}^{N} \exp\{q_k\}$ runs over all $N$ states and is responsible for the probability normalization to unity.

E3: The information about the environment states is updated continuously, which describes the dynamics of the agent’s evaluation of the system state preference (Fig. 1).

Unfortunately, this model cannot be applied directly to describing human intermittent control even within the standard generalizations. The matter is that on time scales much larger than the elementary step $\tau$ in the decision-making process, the time dynamics of state choice is strongly discontinuous. The preference for an option is reflected only in the cumulative number of moments when the agent has selected the given option. To endow human actions with some continuity in time, the desired decision-making process must typically demonstrate the choice of one option on scales comprising many elementary steps of decision-making, what is equivalent to the existence of some decision inertia.

The purpose of the present work is to generalize the reinforcement learning paradigm to make it applicable to modeling human intermittent control. We turn to the concept of status quo bias—preference for the current state of affairs—in human behavior [12], for a recent review of status quo bias see, e.g., [13]. Due to status quo bias the agent selects continuously one option for a relatively long sequence of decision-making events, which endows its actions with quasi-continuous dynamics. It opens a gate to describing human intermittent control as (i) continuous accumulation of information about the state of controlled system within a certain priority function and (ii) the sequence of events when the operator makes decision on changing the control state in response to the accumulated information.

In the present paper we propose a fairly simple model for reinforcement learning with status quo bias and demonstrate that it does meet the desired property of quasi-continuous dynamics. The gist of this model is the concept of multichannel information processing used previously [14] for describing human learning affected by novelty-seeking (intrinsic motivation).

It should be noted that the status quo bias can be seen as a sort of decision inertia implying a perceptual rather than value-based mechanism. Recently [15] a new reinforcement-learning model has been proposed within the classical one-channel paradigm to explain the human decision bias in favor of the same decision as in the previous trial. This model is partly incorporated into the multichannel model to be developed below. Previously [16] we presented the main idea of the developed approach, here the model of reinforcement learning with decision inertia is described in detail.

2 Model

An agent is assumed to make repeated choice between a finite number of options (states of controlled
system) \(i = 1, 2, \ldots, N\) and to accumulate the information about the chosen states. The accumulated information affects its decision-making which is reflected in time variations of the choice probability. We consider the information processing to be implemented through two independent mental channels (Fig. 2). One of them (channel \(Q\)) is the deliberate analysis of the obtained rewards, the other (channel \(\mathcal{A}\)) is irrational and exhibits the status quo bias which can be justified turning to a reason like this: “If just now we have chosen a new option it is not reasonable to choose another option immediately, it could be much better to wait some time until the quality of the chosen option becomes clear.” The two channels interact via their cumulative effect on the option selection (Fig. 2).

The detailed description of the proposed model is as follows. In a regular sequence of time moments \(t_k = k\tau\) \((k \in \mathbb{N} \text{ and } \tau \text{ is the time lag between successive time moments})\) the agent selects one of possible states of controlled system and keeps the system at this state for the time \(\tau\). After that it may select another system state including the previous one. As a result, the agent choice gives rise to system dynamics illustrated in Fig. 3. The “elementary” transition of the system between two successive moments of agent’s choice can be conceived of as:

- at time \(t_{k-1}\) the agent makes decision of transition between the states \(x_i \rightarrow x_j\) based on the current system state;
- then during the time interval \((t_{k-1}, t_k)\) the system kept at the state \(x_j\) evolves according to its own regularities.

These features are described as follows.

### 2.1 Deliberate Information Processing

Via the channel \(Q\) each state \(i\) is related with the corresponding reward \(r_i\), which generally is non-stationary, \(r_i = r_i(t)\), and the agent receives it each time the state \(i\) is chosen. The preference of choosing the state \(i\) is quantified by a value \(q_i\); the value \(q_i\) results from the experience the agent gains each time it chooses the corresponding state.

\[
\Delta_i(t+1) = (1 - \epsilon_q)\Delta_i(t) + \delta_{\mu_k} r_i(t), \quad (1)
\]

where the index \(i_k\) points to the state \(i_k\) chosen at the given time step \(t_k\), the value \(0 < \epsilon < 1\) quantifies the agent memory capacity, and \(\delta_{\mu_k}\) is the Kronecker delta: \(\delta_{\mu_k} = 1\) if \(i = i_k\) and 0 if \(i \neq i_k\). The time scale \(T_q = 1/\epsilon_q\) estimates the duration of agent memory. Expression (1) reflects our assumption that only the preference value of the currently chosen option is increased by the obtained rewards. It is justified by that the effects of “foregone payoffs” may be ignored in the context of the present analysis. For the same reason we do not take into account the fact that people usually overweight low-probability events and underweight high-probability events.

### 2.2 Irrational Information Processing

The channel \(\mathcal{A}\) allows for the effect of status quo bias on selecting the same state at the next time moment of decision-making. This effect is taken into account through some additive preference value \(\Delta_i > 0\) to select the same state \(i\). This value decreases slowly as the time interval of keeping the same state continuously increases and, vice versa, the value \(\Delta_i\) increases with a certain saturation if the given state has not been selected within a certain time interval. Such time variations in the value \(\Delta_i\) are governed by the equation:

\[
\Delta_i(t+1) = (1 - \epsilon_\Delta)\Delta_i(t) + (1 - \delta_{\mu_k}\delta_{\mu_{i,k-1}})\rho_i, \quad (2)
\]

where the coefficient \(\rho_i\) specifies the restoration of the status quo bias measure \(\Delta_i\) for the state \(i\) when it has not been chosen. The given model imitates the human
preference to wait for a certain time $T_a \sim 1/\epsilon_a$ to recognize the quality of the made choice.

2.3 Decision-Making

The proposed model considers that the preference quantities related to the two channels individually meet the following requirements:

- The channels $Q$ and $A$ are assumed to be mutually independent, which implies that the resulting probability of selecting any state $i$ is just the product of their individual contributions.

- The partial probability of selecting a state $i$ is specified by the corresponding priority value.

- The resulting probabilities of selecting states $\{x_i\}$ actually depend only on the difference between the priority values attributed individually to the available options for choice. In other words, their absolute values have no physical meaning. In mathematical terms it implies the invariance of the choice probability with respect to the transformation

$$q_i \rightarrow q_i + C(t).$$

(3)

- The measure of the status quo bias $\Delta_i$ also admits the interpretation as the priority value meeting the previous property. In the given analysis its values is actually fixed by the assumption that the channel $A$ preference of choosing a different state is set equal to zero.

As can be shown, under such conditions the probability of selecting a state $i$ is

$$p_i(q_i) = \frac{1}{Z} \exp \left\{ \beta \left[ q_i + \Delta_i \delta_{ii_{i-1}} \right] \right\}. \tag{4}$$

The normalization coefficient $Z$ is determined by the expression

$$Z = \sum_{i=1}^{N} \exp \left\{ \beta \left[ q_i + \Delta_i \delta_{ii_{i-1}} \right] \right\}. \tag{5}$$

the parameter $\beta$ characterizes the fuzzy threshold in the human cognition of the difference in the priority values between the available options.

2.4 Comment: Random Walk Interpretation

The proposed concept of reinforcement learning with decision inertia may be also regarded as nonlinear (non-Markovian) random walks in the space of system states. Actually the developed model of reinforcement learning admits the representation in terms of walker transitions to another state $j$ provided currently the walker occupies a state $i$. Namely, the probability of this transition can be written as

$$P_{i \rightarrow j} = \frac{1}{Z'} \exp \left\{ \beta \left[ q_j - q_i \right] \right\},$$

and the probability of choosing the same option $i$ is given by the expression

$$P_{i \rightarrow i} = \frac{1}{Z'} \exp \left\{ \beta \Delta_i \right\}. \tag{4}$$

Here the normalization factor $Z'$ now depending on the current state $i$ is related to the previous one, Eq. (5), as

$$Z' = Z \cdot \exp \{\beta q_i \}.$$

The constant value $\beta$ is a system parameter.

3 Simulated Model

The developed concept of reinforcement learning has been analyzed numerically based on the constructed model rewritten in dimensionless form where all the priority values and rewords are measured in units of $\beta^{-1}$. Namely, at each time moment $t_k$ of agent’s decision-making:

* at the first step, the agent selects a state $i$—denoted that as $a_i$—with the probability

$$P_i(t_k) = \frac{\exp \left\{ q_i(t_k) + \Delta_i(t_k) \delta_{ii_{i-1}} \right\}}{\sum_{j=1}^{N} \exp \left\{ q_j(t_k) + \Delta_j(t_k) \delta_{jj_{j-1}} \right\}}; \tag{6a}$$

* then, the preference values of the agent choice are updated according to the expression

$$q_i(t_{k+1}) = (1 - \epsilon_a)q_i(t_k) + \left( \delta_{ii_k} - \frac{1}{N} \right) r_i(t_k); \tag{6b}$$

* and the status quo bias measures are updated as

$$\Delta_i(t_k) = (1 - \epsilon_a)\Delta_i(t_{k-1}) + (1 - \delta_{ii_k} \delta_{ii_{i-1}}) \rho_i. \tag{6c}$$

Here the coefficients

$$\epsilon_q = \frac{1}{T_q}, \quad \epsilon_a = \frac{1}{T_a}, \quad r_i = \frac{q_i^{\max}}{T_q}, \quad \rho_i = \frac{\Delta_i^{\max}}{T_a},$$

are calculated using the time scales $T_q$, $T_a$ characterizing the agent memory decay and the decay of status quo bias, respectively, and the quantities $q_i^{\max}$, $\Delta_i^{\max}$ specifying the saturation of the reword accumulation via the channel $Q$ and the maximum strength of the status-quo-bias effect, respectively.

It should be noted that Eq. (6b) has been obtained using the time function $C(t_k)$ in Exp. (3) such that the identity

$$\sum_{j=1}^{N} q_j(t_k) = 0$$

holds for all the time moments $\{t_k\}$.

The initial condition for all the preference values are set equal to zero and at the initial time moment $t = 0$ all the values of status quo bias are set equal to their maximal values:

$$q_i|_{t=0} = 0 \quad \text{and} \quad \Delta_i|_{t=0} = \Delta_i^{\max} \quad \text{for} \forall i.$$
4 Results: Decision Inertial Effect

In order to verify whether the decision inertia is able to endow the reinforcement learning with the required quasi-continuous dynamics the dimensionless model (6) has been solved numerically. For the sake of simplicity the agent choice has been confined to two options \( i = 1, 2 \) with identical properties, \( r_1 = r_2 = r \) and \( \rho_1 = \rho_2 = \rho \). The main results are shown in Fig. 4.

Figure 4, left column (plots Ia–Va) depicts the results of simulation when the effect of decision inertia (status-quo-bias effect) is weak. In this case the agent behavior is practically unbiased and the reinforcement learning may be considered to be governed solely by the channel \( \mathcal{Q} \). The system parameter were chosen such that the distribution of preference value, i.e., \( P_1(q_1) \) undergo phase transition to the bimodal form. It corresponds to emergence of long-lived states with the choice preference given to one of the two states. It is clear visible in plots Ia and IIIa showing the time patterns of the preference value \( q_1(t) \) (plot Ia) and the status-quo-bias measure \( \Delta_1(t) \) (plot IIIa). In particular, such time patterns imitate a situation when the agent prefers one of the states due to its objective quality and transitions in this preference when the state quality changes with time. These long lived states are characterized by the mean duration about 2000 time units of \( \tau \) when the agent mainly chooses only one state. Unfortunately, the latter statement is justified only when this choice is treated “on the average,” i.e., dealing with its probabilistic properties rather than its continuity in selecting states in the sequence of decision-making events. Indeed, as shown in plot Ia, during a time interval about 1000 units the agent’s choice changes many times. It is also reflected in the distribution of states’ life time \( \tau_r (\tau_{ile}) \) depicted in plot IVa. Here the life time \( \tau_{ile} \) of a state is defined as the duration of time interval during which the agent chooses this state continuously. The found mean value \( \langle \tau_{ile} \rangle \) turns out to be very small, \( \langle \tau_{ile} \rangle \approx 5 \), in comparison of the mean duration \( \approx 2000 \) of long lived states. The autocorrelation function of selected states (plot Va) also demonstrates this feature. Finalizing the presented analysis of the reinforcement learning with weak status-quo-bias effect (or its absence) we may draw the conclusion that the standard paradigm of reinforcement learning cannot be applied directly to modeling human intermittent control.

Figure 4, right column (plots Ib–Vb) depicts similar results when the decision inertia affects the agent behavior substantially. It should be emphasized that the results of simulation presented on the left and right columns were obtained when the analyzed systems differ only in the status-quo-bias measure (see the caption to Fig. 4). In plots Ib–IIIb the shown time patterns of the states’ selection \( S(t) \) (plot Ib), the preference value \( q_1(t) \) (plot IIb), and the status-quo-bias measure \( \Delta_1(t) \) demonstrate rather regular behavior with continuous (or smooth) fragments of duration about 50 time units of \( \tau \). It is also justified by the found distribution of states’ life time (plot IVb) and the corresponding autocorrelation function of selected states (plot Vb). Moreover, as seen in plot IVb, states with short life time about several units of \( \tau \) are really rare and, so, may be ignored. Besides, these time patterns correspond to the agent behavior when it rather regularly selects different states for analyzing their quality within time scales much larger than the time leg \( \tau \) between successive events of decision-making. Exactly the latter feature demonstrates the applicability of initially discrete model for describing human actions with inherently quasi-continuous properties. It opens a gate to applying the formalism of reinforcement learning with decision inertia to modeling human intermittent control.

5 Conclusion

In the present work we have posed a question about the possibility of constructing a theory of human intermittent control treated a sequence of events when the operator continuously makes decision about activating or halting the control. These decision events are based on the continuous processing of information about the dynamics of controlled system and estimating the preference of active or passive behavior in governing the system dynamics. In the standard terms using in describing human intermittent control with event driven activation as an alternate sequence of active and passive phases of human actions the description we have proposed takes into account:

- the probabilistic nature of decision-making in activating or halting the control over an unstable system, which corresponds to the concept of noise-induced activation [9, 10];
- the interpretation of the passive phase fragments as the moments when the operator accumulates the information about the state of controlled system with respect to its deviation from the desired state;
- the interpretation of the active phase fragments as the open-loop control when the operator mainly implements planned actions rather than responds to local changes in the system motion;
- the probabilistic transitions between the active and passive phases as timeless events;

We have put forward the formalism of reinforcement learning as a prospective candidate for describing human intermittent control with such properties. This formalism provides a rather natural framework of copying with decision-making in response to accumulated information about the controlled system via the trial-error strategy. Unfortunately, as demonstrated,
Fig. 4: Time patterns and statistical properties of reinforcement learning for the system with two equivalent states. In simulation the following parameters were used: $T_q = T_a = 30$, $q^{\text{max}} = 2.5$, and $\Delta^0_1 = \Delta^0_2 = 1$ or $\Delta^0_1 = \Delta^0_2 = 10$ for the cases of for the case of weak and strong status-quo-bias effect, respectively. Duration of simulation was $10^6$ time units of $\tau$. 
the standard formalism of reinforcement learning cannot provide the description with the required quasi-continuous behavior of the operator in selecting the appropriate system states.

To overcome this problem we have proposed a novel generalized formalism of reinforcement learning with decision inertia. The gist of this formalism is the introduction of two independent channels of information processing:

- the channel $Q$ responsible for the deliberate analysis of the obtained rewards and ordering them according to preference; it is the standard component of all the models of reinforcement learning;

- the channel $A$ of irrational information processing that stimulates the operator not to change the current choice and keep it for a certain time interval; it is called the decision inertia.

Actually the two channels interacting with each other when the operator makes a decision endows the reinforcement learning with multidimensional dynamics and complex properties.

As demonstrated, the formalism of reinforcement learning with decision inertia is able to simulate quasi-continuous dynamics of agent’s choice, which opens a gate toward the desired description of human intermittent control.

References

[1] P. Gawthrop, I. Loram, M. Lakie, and H. Gollee. Intermittent control: a computational theory of human control. *Biol. Cybernetics*, 104:31–51, 2011.

[2] I. D. Loram, H. Gollee, M. Lakie, and P. J. Gawthrop. Human control of an inverted pendulum: is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *J. Physiology*, 589(2):307–324, 2011.

[3] R. Balasubramaniam. On the Control of Unstable Objects: The Dynamics of Human Stick Balancing. In M. J. Richardson, M. A. Riley, and K. Shockley, editors, *Progress in Motor Control: Neural, Computational and Dynamic Approaches*, pages 149–168. Springer, New York, 2013.

[4] J. G. Milton. Intermittent Motor Control: The “drift-and-act” Hypothesis. In M. J. Richardson, M. A. Riley, and K. Shockley, editors, *Progress in Motor Control: Control Neural, Computational and Dynamic Approaches*, pages 169–193. Springer, New York, 2013.

[5] Y. Asai, S. Tateyama, and T. Nomura. Learning an Intermittent Control Strategy for Postural Balancing Using an EMG-Based Human-Computer Interface. *PLoS ONE*, 8(5):e62956, 2013.

[6] J. Milton, T. Insperger, and G. Stepan. Human Balance Control: Dead Zones, Intermittency, and Micro-chaos. In Toru Ohira and Tohru Uzawa, editors, *Mathematical Approaches to Biological Systems: Networks, Oscillations, and Collective Motions*, pages 1–28. Springer Japan, Tokyo, 2015.

[7] Y. Asai, Y. Tasaka, K. Nomura, T. Nomura, M. Casadio, and P. Morasso. A model of postural control in quiet standing: robust compensation of delay-induced instability using intermittent activation of feedback control. *PLoS One*, 4(7):e6169 (1–14), 2009.

[8] N. Yoshikawa, Y. Suzuki, K. Kiyono, and T. Nomura. Intermittent Feedback-Control Strategy for Stabilizing Inverted Pendulum on Manually Controlled Cart as Analogy to Human Stick Balancing. *Frontiers in Computational Neuroscience*, 10(Article 34):19 pages, 2016.

[9] A. Zgonnikov, I. Lubashevsky, S. Kanemoto, T. Miyazawa, and T. Suzuki. To react or not to react? Intrinsic stochasticity of human control in virtual stick balancing. *Journal of The Royal Society Interface*, 11:20140636, 2014.

[10] A. Zgonnikov and I. Lubashevsky. Double-well dynamics of noise-driven control activation in human intermittent control: the case of stick balancing. *Cognitive Processing*, 16(4):351–358, 2015.

[11] M. P. Deisenroth, G. Neumann, J. Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

[12] W. Samuelson and R. Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, 1988.

[13] S. Eidelman and C. S. Crandall. Bias in favor of the status quo. *Social and Personality Psychology Compass*, 6(3):270–281, 2012.

[14] A. Zgonnikov and I. Lubashevsky. Unstable Dynamics of Adaptation in Unknown Environment due to Novelty Seeking. *Advances in Complex Systems*, 17(3 & 4):1450013 (17 pages), 2014.

[15] R. Akaishi, K. Umeda, A. Nagase, and K. Sakai. Autonomous Mechanism of Internal Choice Estimate Underlies Decision Inertia. *Neuron*, 81(1):195–206, 2014.

[16] K. Hijiwata and I. Lubashevsky. Reinforcement Learning with Status Quo Bias. In Y. Yamamoto, T. Nomura, S. Cerutti, H. Dickhaus, and K. Yama, editors, *The 8th International Workshop on Biosignal Interpretation (BSI2016), Proceedings of (Osaka, Nov. 1–3)*, pages 201–204, 2016.