Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 12 (2011) 164–174

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Psychological models of human and optimal performance in bandit problems

Action editor: Andrew Howes

Michael D. Lee [*], Shunan Zhang, Miles Munro, Mark Steyvers

Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA

Received 7 December 2009; accepted 9 July 2010
Available online 11 August 2010

Abstract

In bandit problems, a decision-maker must choose between a set of alternatives, each of which has a fixed but unknown rate of reward, to maximize their total number of rewards over a sequence of trials. Performing well in these problems requires balancing the need to search for highly-rewarding alternatives, with the need to capitalize on those alternatives already known to be reasonably good. Consistent with this motivation, we develop a new psychological model that relies on switching between latent *exploration* and *exploitation* states. We test the model over a range of two-alternative bandit problems, against both human and optimal decision-making data, comparing it to benchmark models from the reinforcement learning literature. By making inferences about the latent states from optimal decision-making behavior, we characterize how people should switch between exploration and exploitation. By making inferences from human data, we begin to characterize how people actually do switch. We discuss the implications of these findings for understanding and measuring the competing demands of exploration and exploitation in sequential decision-making.

Keywords: Bandit problem; Exploration versus exploitation; Heuristic models; Latent state models; Reinforcement learning

1. Introduction

1.1. Bandit problems

In bandit problems, a decision-maker chooses repeatedly between a set of alternatives. They get feedback after every decision, either recording a reward or a failure. They also know that each alternative has some fixed, but unknown, probability of providing a reward each time it is chosen. The goal of the decision-maker is to obtain the maximum number of rewards over all the trials they complete. In some bandit problems, the number of trials is not known in advance, but there is some probability any trial will be the last. These are known as ‘infinite horizon’ bandit problems. In other bandit problems the number of trials is fixed, known, and usually small. These are known as ‘finite-horizon’ bandit problems.

Bandit problems provide an interesting formal setting for studying the balance between exploration and exploitation in decision-making. In early trials, it makes sense to explore different alternatives, searching for those with the highest reward rates. In later trials, it makes sense to exploit those alternatives known to be good. How exactly this balance between exploration and exploitation should be managed, and should be influenced by factors such as the distribution of reward rates, the total number of trials, and so on, raises basic questions about adaptation, planning, and learning in intelligent systems. For these reasons, bandit problems have been widely studied in machine learning (Berry & Fristedt, 1985; Gittins, 1979; Kaebling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1998) and cognitive science (Cohen, McClure, & Yu, 2007; Daw, O’Doherty,

[*] Corresponding author.
E-mail address: mdlee@uci.edu (M.D. Lee).

1389-0417/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2010.07.007

Dayan, Seymour, & Dolan, 2006; Steyvers, Lee, & Wagenmakers, 2009), and many models of decision-making strategies have been proposed.

1.2. Research goals

A first motivation for our work is to refine and extend one existing theoretical idea that seems especially relevant to understanding human decision-making on bandit problems. This is the idea of latent state modeling, in which behavior is treated as a mixture of different processes, controlled by unobserved states. Latent state models are well suited to situations, where two or more qualitatively different types of decision-making are needed to explain performance as a whole. The general latent state approach has been successful in many areas of the cognitive sciences, ranging from all-or-none theories of learning (Batchelder, 1970), to models of language (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), to models of the roles of guessing and other contaminant behavior in simple decision-making (Vandekerckhove & Tuerlinckx, 2007). The latent state approach seems a particularly natural account of human decision-making in bandit problems, given the requirement to choose between the competing, and qualitatively different, demands of exploration and exploitation.

A second motivating challenge for our work involves interpreting, evaluating and potentially improving human decision-making. Using the optimal decision process (Kaebling et al., 1996), it is possible to evaluate how well a person solves bandit problems. The conclusion might be something like “you got 67% rewards, but optimal behavior would have given you 75% rewards, so you are falling short.” This seems like only a partial evaluation, because it does not explain *why* their decisions were sub-optimal. Instead, to help us understand human and optimal decision-making on bandit problems, we use simple heuristic models. These include several benchmark models from the existing machine learning literature, as well the new latent state model we develop. The attraction of these models is that they provide simple process accounts of how a decision-maker should behave, depending on a small set of parameters. We choose models whose parameters have clear and useful psychological interpretations. This means that, when we fit the models to data, and estimate the parameters, we obtain interpretable measures of key aspects of decision-making. Instead of just telling people they are falling short of optimal, we now aim also to tell them “the problem seems to be you are exploring for too long: the optimal thing to do is to stop exploring at about the 5th trial”, or “you are not shifting away quickly enough from a choice that is failing to reward you: the optimal thing to do is to leave a failed choice about 80% of the time.”

1.3. Overview

With these motivations in place, the outline of this paper is as follows. First, we describe an experiment in which human and optimal decision-making data for a variety of bandit problems was collected. We then describe four existing benchmark heuristics, before developing our new model. We test all of these models as accounts of the human and optimal decision data, using Bayesian methods that balance both goodness-of-fit and model complexity in model evaluation, and find that our new model performs better than the existing ones. Finally, we demonstrate how the psychological interpretability of the heuristics can help characterize and compare human and optimal decision-making on bandit problems.

2. Human and optimal decision data

2.1. Bandit problem conditions

We considered six different types of bandit problems, all involving just two-alternatives, which is the most commonly studied case in the literature, and all having short fixed horizons. The six conditions varied in a 2×3 design, manipulating how many trials there were in a problem, and how the reward rates for the alternatives were chosen. Specifically, there were two trial sizes (8-trial and 16-trial), and three different environmental distributions (‘plentiful’, ‘neutral’ and ‘scarce’) controlling the reward rates.

The basic idea of environmental distributions is to manipulate whether reward rates tend to have high or low values. Following (Steyvers et al., 2009), the environments were defined in terms of Beta (α, β) distributions, where α corresponds to a count of ‘prior successes’ and β to a count of ‘prior failures’. The plentiful, neutral and scarce environments used, respectively, the values $\alpha = 4$, $\beta = 2$, $\alpha = \beta = 1$, and $\alpha = 2$, $\beta = 4$. Reward rates for each alternative in each problem were sampled independently, for a total of 50 problems in each condition, from the appropriate environmental distribution.

2.2. Human data

Data were collected from 10 naive participants (6 males, 4 females). A representation of the basic experimental interface is shown in Fig. 1. The two large panels correspond to the alternatives, either of which can be chosen on any trial by pressing the button below. Within the panel, the outcomes of previous choices are shown as count bars, with successes on the left, and failures on the right. At the top of each panel, the proportion of successes, if defined, is shown. The top of the interface provides the success count, the current trial number, the total number of trials, and a count of how many problems out of the entire set have been completed.

Using this interface, within-participant data were collect for all 50 problems for all six bandit problem conditions. The order of the conditions, and of the problems within the conditions, was randomized for each participant. All $6 \times 50 = 300$ problems (as well as five practice problems per condition) were completed in a single experimental session, with breaks taken between conditions.