MULTI-ARMED BANDITS FOR HUMAN-MACHINE DECISION MAKING

Paul Reverdy

University of Arizona
Aerospace and Mechanical Engineering
Tucson, AZ 85719

Vaibhav Srivastava[*]

Michigan State University
Electrical and Computer Engineering
East Lansing, MI 48824

ABSTRACT

Building an integrated human-machine decision-making system requires developing effective interfaces between the human and the machine. We develop such an interface by studying the multi-armed bandit problem, a simple sequential decision-making paradigm that can model a variety of tasks. We construct Bayesian algorithms for the multi-armed bandit problem, prove conditions under which these algorithms achieve good performance, and empirically show that, with appropriate priors, these algorithms effectively model human choice behavior; the priors then form a principled interface from human to machine. We take a signal processing perspective on the prior estimation problem and develop methods to estimate the priors given human choice data.

Index Terms— Active inference, Bayesian inference, multi-armed bandit, human decision making

1. INTRODUCTION

Inference, the process of reaching conclusions from data, lies at the heart of many contemporary technologies, including object recognition and fault detection among numerous others. Often, such technologies are employed not directly for inference but rather to infer some information (the type of a perceived object, the presence of a fault, etc.) in order to take an action (manipulate the perceived object, isolate and recover from the fault, etc.). In this paper, we study the multi-armed bandit (MAB) problem as a simple example of a decision-making task where inference and action are closely linked.

Additionally, many inference problems are highly structured in the sense that only small amounts of data are required to perform accurate inference; often this is due to the existence of a great deal of contextual information, such as the types of objects or faults that the system is likely to encounter. Providing this contextual information can greatly improve the performance of the system, e.g., its convergence rates, it is of great interest to develop methods to do so. Contextual information is often difficult to systematically extract, so many deployed systems use human supervisors to provide the contextual information and to guide the automated system [1, 2]. The performance of such human-machine systems could be improved by rigorously studying the connections between human and machine decision making and thereby developing effective interfaces between the human and the machine.

In this paper, we consider the MAB problem from the viewpoint of both the human and the machine and develop the Upper Credible Limit (UCL) algorithm, a Bayesian algorithm that models the principal features of human choice behavior in MAB problems while also achieving optimal performance when employed with appropriate values of its input parameters, principally the algorithm's priors. These input parameters then form a parsimonious quantitative interface between human and machine, and we develop methods to estimate the priors given human choice data. Significant portions of these results have previously appeared in various publications, including [3, 4, 5], and [6].

2. THE MULTI-ARMED BANDIT (MAB) PROBLEM

The multi-armed bandit (MAB) problem, introduced by Robbins [7], is a sequential decision-making problem in which a decision-making agent is presented with a set of N options (an option is also called an *arm* in analogy with the lever of a slot machine). Each option $i \in \{1, \dots, N\}$ has an associated probability distribution p_i whose mean m_i is unknown to the decision maker. At each sequential decision time $t \in \{1, \dots, T\}$ the agent picks arm i_t and receives reward $r_t \sim p_{i_t}(r)$ drawn from the probability distribution associated with arm i_t. The agent's objective is to pick arms such that the expected value of the rewards received from the T decisions is maximized:

$$\max_{\{i_t\}} J, \quad J = \mathbb{E}\left[\sum_{t=1}^{T} r_{i_t}\right] = \sum_{t=1}^{T} \mathbb{E}\left[m_{i_t}\right], \tag{1}$$

where the latter expectation is over different realizations of the sequence $\{i_t\}$.

Each choice of i_t is made sequentially, conditional on the information available to the agent at time t. If the mean rewards m_i were known to the agent *a priori*, the optimal policy would be trivial: set $i_t = \arg\max_i m_i$ for each t. However,

[*]This work has been supported by NSF Award IIS-1734272.

since the mean rewards are not known, the agent must simultaneously seek to select arms for which the rewards are poorly known (explore) and select arms that appear to have high rewards based on current information (exploit). The tension between selecting arms with uncertain (but potentially high) rewards and arms that appear to have high rewards is known as the *explore-exploit tradeoff*, and is common to problems in active learning and adaptive control. In the literature, the rewards are often assumed to be Bernoulli and model, e.g., whether or not an individual will click on an ad on a website [8], however other reward distributions have been considered.

2.1. Performance bounds for MAB problems

In the MAB literature, it is common to assume that the reward distributions p_i are stationary. Under this assumption, a famous result due to Lai and Robbins [9] bounds the performance of any algorithm solving the MAB problem. The bound is typically stated in terms of *regret*, which is a measure of performance loss due to uncertainty. Defining $m_{i^*} = \max_i m_i$ and $R_t = m_{i^*} - m_{i_t}$ as the *expected regret* (conditioned on i_t) at time t, the objective (1) can be rewritten as minimizing the *cumulative expected regret* defined as

$$\sum_{t=1}^{T} R_t = T m_{i^*} - \sum_{t=1}^{T} m_{i_t} = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[n_i^T\right],$$

where n_i^T is the number of times arm i has been chosen up to time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to choosing arm i instead of i^*. To minimize the cumulative expected regret it suffices to minimize the number of times a suboptimal arm $i \in \{1, \ldots, N\} \backslash \{i^*\}$ is selected.

Lai and Robbins [9] showed that the expected number of times a suboptimal arm is selected is at least logarithmic in time, i.e.,

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{1}{D(p_i||p_{i^*})} + o(1)\right) \log T \tag{2}$$

for each $i \in \{1, \ldots, N\} \backslash \{i^*\}$, where $o(1) \to 0$ as $T \to +\infty$ and $D(p_i||p_{i^*})$ is the Kullback-Leibler divergence between p_i and p_{i^*}. This bound implies that the cumulative expected regret must grow at least logarithmically with time. In the literature, algorithms that achieve cumulative expected regret that is uniformly bounded by a logarithmic term with a constant that is within a constant factor of (2) are said to achieve *logarithmic regret* and considered to have optimal performance.

In the remainder of this paper, we focus on the case of Gaussian rewards, i.e., where p_i is Gaussian with mean m_i which is unknown to the decision maker and variance $\sigma_{s,i}^2$, which is known. If, in addition, the reward variance is uniform (i.e., $\sigma_{s,i} = \sigma_s$), then the constant $1/D(p_i||p_{i^*})$ in (2) reduces to $2\sigma_s^2/\Delta_i^2$.

2.2. Features of human decision-making behavior in MAB problems

In [3], we identified five salient features of human decision-making behavior in MAB problems. These features are likely to be apparent in human decision-making behavior in other problems as well so we repeat them here as follows.

1. Familiarity with the environment: Humans approach problems with prior knowledge, which here is manifest as prior knowledge about the mean reward m_i associated with each arm i.

2. Ambiguity bonus: Wilson *et al.* [10] have shown that human decision-making in MAB problems is based on a linear combination of an estimate of the mean reward m_i and the uncertainty in that estimate.

3. Stochasticity: Human decision-making behavior is inherently noisy [11].

4. Finite-horizon effects: Both the level of decision noise and the ambiguity bonus effect are sensitive to the time horizon T [10].

5. Environmental structure effects: Humans tend to learn the structure of the tasks they perform, i.e., they learn the correlation structure among the rewards from different arms, and use this structural information to improve their decisions [12].

3. UCL: A BAYESIAN MAB ALGORITHM

In [3], we developed an algorithm called the Upper Credible Limit (UCL) algorithm for solving MAB problems with Gaussian rewards. UCL is a Bayesian algorithm inspired by the Bayes-UCB algorithm developed by Kauffman *et al.* [13] for the case of Bernoulli rewards. Both algorithms are based on the *optimism in the face of uncertainty* heuristic [8], which suggests that algorithms can achieve good performance by formulating the set of possible environments (i.e., reward distributions) that are consistent with observed data, then acting as if the true environment were a sufficiently favorable one in that set.

UCL maintains a belief state about the rewards using Bayesian inference. Let $m \in \mathbb{R}^N$ be the vector of unknown mean rewards. We let the prior on m be the Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, where $\mu_0 \in \mathbb{R}^N$ and $\Sigma_0 \in \mathbb{R}^{N \times N}$ is a positive-definite matrix. Since the reward distributions are assumed to be Gaussian with known variance $\sigma_{s,i}^2$, the Gaussian prior is conjugate to the observation likelihood and the belief state (μ_t, Σ_t) updates in closed form according to well-known linear equations [14]. The marginal distribution for the i^{th} component of the belief state is then the Gaussian distribution $\mathcal{N}(\mu_i^t, \sigma_i^t)$, where $\mu_i^t = (\mu_t)_i$ and $\sigma_i^t = \sqrt{(\Sigma_t)_{ii}}$.

3.1. The deterministic UCL algorithm

At each decision time $t \in \{1, \ldots, T\}$, the deterministic UCL algorithm selects the arm that maximizes the upper limit of the $(1 - 1/Kt)$ credible interval, i.e., it selects an arm $i_t = \arg\max_i Q_i^t$, where

$$Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - 1/Kt),$$

$\Phi^{-1} : (0, 1) \to \mathbb{R}$ is the inverse cdf of the standard Gaussian random variable, and $K \in \mathbb{R}_{>0}$ is a tunable parameter.

3.2. The stochastic UCL algorithm

As noted as feature 3 in Section 2.2, human decision-making behavior in MAB problems is inherently noisy. Therefore, in [3], we considered an extension of the deterministic UCL algorithm which models the decision noise using the Boltzmann selection mechanism, i.e., sets the probability p_i^t of picking arm i at time t equal to

$$p_i^t = \frac{\exp(Q_i^t/\nu_t)}{\sum_{j=1}^N \exp(Q_j^t/\nu_t)}.$$

In the limit $\nu_t \to 0^+$, this scheme chooses $i_t = \arg\max_i Q_i^t$ and as ν_t increases the probability of selecting any other arm increases. Thus, Boltzmann selection generalizes the maximum operation and is sometimes called the soft maximum (or softmax) rule.

The parameter ν_t is known as the temperature parameter of the softmax, and the functional form of the parameter ν_t is known as a *cooling schedule*. In [3], we showed that cooling schedules of the form $\nu_t = \nu/\log t, \nu > 0$ are effective in modeling human behavior.

3.3. Performance guarantees for UCL

In [3], we studied the case of homogenous sampling noise (i.e., $\sigma_{s,i}^2 = \sigma_s^2$ for each i) and showed that the UCL algorithm achieves logarithmic regret with an uncorrelated uninformative prior. Specifically, we proved the following theorem.

Theorem 1 (Regret of the Deterministic UCL algorithm)**.** *The following statements hold for the Gaussian multi-armed bandit problem and the deterministic UCL algorithm with uncorrelated uninformative prior and $K = 1$:*

1. *the expected number of times a suboptimal arm i is chosen until time T satisfies*

$$\mathbb{E}\left[n_i^T\right] \leq \left(\frac{8\sigma_s^2}{\Delta_i^2} + 2\right) \log T + 3;$$

2. *the cumulative expected regret until time T satisfies*

$$J_R = \sum_{i=1}^T R_t \leq \sum_{i=1}^N \Delta_i \left(\left(\frac{8\sigma_s^2}{\Delta_i^2} + 2\right) \log T + 3\right).$$

The implication of this theorem can be seen by comparing the first statement of the theorem with the Lai-Robbins bound (2): the deterministic UCL algorithm achieves logarithmic regret and thus is considered to have optimal performance. A similar theorem holds for the stochastic UCL algorithm with appropriate tuning rule for ν; see [3, Theorem 7].

3.4. UCL as a model of human decision making

In [3], we studied data from a human-subject study where participants solved a spatially-embedded MAB problem. By spatially-embedded MAB problem, we mean a MAB problem where the arms correspond to patches of space, in this case, squares in a 10×10 grid. Such a problem might model, e.g., function optimization in a discretized space. The spatially-embedded MAB problem then inherits structure from the smoothness of the function being optimized.

We showed that, in this problem, subject performance largely fell into two categories which we termed *phenotypes*: one corresponded to cumulative regret depending approximately linearly on T and represented poor performance, while the other corresponded to cumulative regret depending approximately logarithmically on T and represented good performance. Subjects displaying the logarithmic phenotype in fact achieved performance better than an otherwise "optimal" algorithm over the short time horizon of the experimental problem, which we ascribed to the subjects' understanding of the structure of the problem. In the spatially-embedded MAB problem, the smoothness of the underlying function being optimized means that the rewards from one arm will be highly correlated with the rewards from nearby arms and less correlated with arms that are farther away.

The assumption of spatial smoothness in mean rewards can be translated into an assumption on the correlation structure of m by choosing the elements of Σ_0 to have the form of an exponential radial basis function with length scale $\lambda \geq 0$ representing the spatial smoothness and overall scale factor $\sigma_0 \geq 0$ representing the strength of the subject's beliefs:

$$(\Sigma_0)_{ij} = \sigma_0^2 \exp(-r_{ij}/\lambda),$$

where r_{ij} is the distance between arms i and j.

Similarly, one can parametrize the prior mean $\mu_0 \in \mathbb{R}^N$ by setting $(\mu_0)_i = \mu \in \mathbb{R}$ for each element i. Finally, by adopting the cooling schedule $\nu_t = \nu/\log t, \nu > 0$, we can parametrize the subject's decision noise with a single parameter. This yields set of parameters consisting of the four scalars $\mu \in \mathbb{R}, \sigma_0 \geq 0, \lambda \geq 0$, and $\nu \geq 0$. In [3], we showed that careful choices of these four parameters allowed the stochastic UCL algorithm to produce behavior that qualitatively matched the two behavioral phenotypes.

4. MODEL FITTING TO HUMAN CHOICE DATA

In [3], we showed that the four-dimensional parameter vector $\theta = (\mu, \sigma_0, \lambda, \nu)$ was sufficient to allow the stochastic UCL algorithm to qualitatively fit human subject data. Therefore, measuring θ would allow a system to quantify a human operator's intuition about the structure of the problem, as expressed through their choice behavior. In [4], we studied the problem of estimating θ from behavioral data in detail using a maximum likelihood approach. We showed that the stochastic UCL algorithm defines a likelihood function in a straightforward way and that this likelihood function can usefully be approximated as a linear function of θ by linearizing about a nominal parameter vector θ_0. We derived conditions under which the resulting maximum likelihood problem is convex and showed that it could be solved by standard convex optimization tools.

We then applied the estimation procedure to data from the human subject experiment first studied in [3]. The experiment design was such that subjects solved one of two tasks, each with a different underlying reward structure, which we referred to as a *landscape*. The subjects presented with each landscape separated into two phenotypes as in [3], resulting in four groups of subjects: high and low performance phenotypes for each of the two landscapes. We used the estimation procedure to estimate θ for each subject and then compared across the four groups of subjects.

The parameter estimates were relatively precise (i.e., small confidence intervals) for the two high-performance groups and relatively imprecise (i.e., large confidence intervals) for the two low-performance groups. This is unsurprising, as the stochastic UCL algorithm is designed to have high performance and it is likely that the set of parameter values corresponding to low performance is large, while the set of values corresponding to high performance is likely to be relatively small. Comparing the parameter estimates across the four groups then showed that there was a statistically-significant difference in parameter values between the two high performance groups but no other statistically-significant differences between groups. One can then conclude that, at least as quantified by θ in the stochastic UCL model, the subjects who had high performance had detectable differences in their intuition about the problem. Therefore, estimating θ from a subject with high performance can provide quantified information about how to achieve high performance with the stochastic UCL algorithm.

5. IMPLICATIONS FOR HUMAN-MACHINE INFERENCE NETWORKS

This work has implications for human-machine inference networks. For applications where inference is being performed to support a decision, it may be possible to model the joint inference-decision task using the multi-armed bandit framework. For cases where the MAB model is appropriate, our findings show that humans can achieve performance that greatly exceeds that of an otherwise "optimal" algorithm, particularly when the time horizon T is short and the MAB problem has significant structure. Furthermore, the stochastic UCL algorithm can be used as a model to quantify the intuition of a human decision maker in terms of its parameter vector θ.

The work in [4] shows that standard maximum likelihood parameter estimation techniques can be used to estimate θ and that statistically-significant information can be extracted from the observed behavioral choice data of high-performing individuals. One could then build a human-machine decision-making system where an expert would perform a certain number of initial choices. These initial choices would be used to estimate the expert's value of θ, and then the system could make autonomous choices by employing the stochastic UCL algorithm the estimated value of θ as input parameters. In this way, the model and estimator could be used to train the stochastic UCL algorithm to make better decisions.

There are a number of clear directions for future research along these lines. One direction is to pursue modeling inference-decision problems in the MAB framework. Gai *et al.* [15] have presented one example from the domain of wireless networking, but there are undoubtedly others. Another direction pertains to the construction of human-machine systems. In this setting, open questions abound: What is the appropriate amount of training for the machine system? When should the machine ask for additional training from the human, for example if it detects that the problem has changed in some way? If the human needs to make the ultimate decision, what information should the machine provide, and in what form (e.g., a suggested action or set of actions)? Undoubtedly the answers to these questions will depend upon the application context, but there are likely fundamental problems to be solved as well.

6. CONCLUSION

In conclusion, we argue that the multi-armed bandit problem is a useful framework for studying the interaction between human and machine in the context of active inference problems. The stochastic UCL algorithm can be considered both as a parametric model of human decision-making behavior in MAB problems and as a computationally-efficient and high-performance method to solve such problems. The algorithm parameters θ form a quantitative, principled interface between human and machine. This framework will allow the construction of joint human-machine inference systems which will raise additional fundamental questions in turn.

7. REFERENCES

[1] S. Lohr, "Algorithms get a human hand in steering web," *The New York Times*, vol. March 11, pp. A1, 2013.

[2] T. O'Reilly, "#IoTH: The internet of things and humans," April 2014, http://radar.oreilly.com/2014/04/ioth-the-internet-of-things-and-humans.html.

[3] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized multi-armed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[4] P. Reverdy and N. E. Leonard, "Parameter estimation in softmax decision-making models with linear objective functions," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 54–67, 2016.

[5] V. Srivastava, P. Reverdy, and N. E. Leonard, "Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis," *arXiv preprint arXiv:1507.01160*, 2015.

[6] P. Reverdy, V. Srivastava, and N. E. Leonard, "Satisficing in multi-armed bandit problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3788–3803, Aug 2017.

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

[8] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[10] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, "Humans use directed and random exploration to solve the explore-exploit dilemma," *Journal of Experimental Psychology: General*, vol. 143, no. 6, pp. 2074–2081, 2014.

[11] M. Steyvers, M. D. Lee, and E. Wagenmakers, "A Bayesian analysis of human decision-making on bandit problems," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 168–179, 2009.

[12] D. E. Acuña and P. Schrater, "Structure learning in human sequential decision-making," *PLoS computational biology*, vol. 6, no. 12, pp. e1001003, 2010.

[13] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *International Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands, Spain, Apr. 2012, pp. 592–600.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I : Estimation Theory*, Prentice Hall, 1993.

[15] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*. IEEE, 2010, pp. 1–9.