Modeling Human Decision-making in Multi-armed Bandits

Paul Reverdy
Department of Mechanical and Aero. Engineering
Princeton University
Princeton, NJ 08544
preverdy@princeton.edu

Vaibhav Srivastava
Department of Mechanical and Aero. Engineering
Princeton University
Princeton, NJ 08544
vaibhavs@princeton.edu

Naomi E. Leonard
Department of Mechanical and Aero. Engineering
Princeton University
Princeton, NJ 08544
naomi@princeton.edu

Abstract

We study the exploration-exploitation trade-off in human decision-making in the context of multi-armed bandit problems. We consider a Bayesian multi-armed bandit problem with Gaussian rewards and develop an efficient algorithm that captures the empirically observed trends in human-decision making. In particular, the proposed algorithm captures the following features observed in human decision-making: (i) increased exploration with increasing time horizon of the decision task, (ii) ambiguity bonus, and (iii) inherent decision-noise. We characterize the efficiency of the algorithm in terms of the regret associated with the decision process. For the no decision-noise case, we demonstrate that as the model parameters encoding the prior knowledge of the human are varied, the performance may change from efficient (logarithmic regret) to the worst case (linear regret).

Keywords: exploration-exploitation trade-off, multi-armed bandits, Bayes-UCB algorithm, human decision-making.

Acknowledgements

This research has been supported in part by AFOSR grant FA9550-07-1-0-0528 and ONR grant N00014-09-1-1074. P. Reverdy is supported through an NDSEG Fellowship.

1 Multi-armed Bandit Problem

Multi-arm bandit problems are a class of resource allocation problems in which a decision-maker allocates a single resource by sequentially choosing one among a set of competing alternative options called arms. In the so-called stationary multi-armed bandit problem, a decision-maker at each discrete time instant chooses an arm and collects a reward drawn from an unknown stationary probability distribution associated with the selected arm. The objective of the decision-maker is to maximize the total reward aggregated over the sequential allocation process. These problems capture the fundamental trade-off between exploration (collecting more information to reduce uncertainty) and exploitation (using the current information to maximize the immediate reward).

The multi-armed bandit problem originated from the analysis of clinical trials [13], where the decision-maker was a doctor and the options were different treatments for a given disease. The doctor's goal in the trial was to find the most efficient treatment while curing as many patients as possible. Robbins [10] later posed the multi-armed bandit problem in the context of sequential design of experiments. The optimal solution to a multi-armed bandit problem can be obtained though stochastic dynamic programming [7] but it quickly becomes intractable as horizon length grows. In his seminal work, Gittins [6] developed a dynamic allocation index for each arm and showed that selecting an arm with the highest index at the given time results in the optimal policy. The dynamic allocation index, while a powerful idea, suffers from two drawbacks: (i) the dynamic allocation index is hard to compute, and (ii) it does not provide insight into the nature of the optimal policies. In another ground-breaking work, Lai and Robbins [9] established a logarithmic lower bound on the expected number of times a sub-optimal arm needs to be sampled by an optimal policy. They also developed an upper confidence bound-based algorithm that achieves the lower bound asymptotically. The computation of the upper confidence bounds in [9] involves tedious computations and Agarwal [3] simplified these computations to develop sample mean-based upper bounds that achieve logarithmic regret. Auer *et al.* [4], in their seminal work, developed upper confidence bound-based algorithms that achieve logarithmic regret uniformly in time. Recently, Srinivas *et al.* [11] developed asymptotically optimal upper confidence bound-based algorithms for Gaussian process optimization. Kauffman *et al.* [8] developed a generic Bayesian upper confidence bound-based algorithm and established its optimality for binary bandits with uniform prior. We develop a similar Bayesian upper confidence bound-based algorithm for Gaussian multi-armed bandit problems and show that it achieves logarithmic regret for uninformative priors. We draw connections between the proposed algorithm and human decision-making in multi-armed bandit problems and highlight the effect of priors on the performance of the proposed algorithm.

2 Gaussian Multi-armed Bandit Problem

Consider an N-armed bandit problem, i.e., a multi-armed bandit problem with N arms. The reward associated with arm $i \in \{1, \ldots, N\}$ is a Gaussian random variable with unknown mean m_i and known variance σ_s^2. The mean of the Gaussian reward at arm i can be interpreted as the signal strength at the arm, while the variance can be interpreted as the sampling noise that is the same at each arm. Let the decision-maker choose arm i_t at time $t \in \{1, \ldots, T\}$ and receive a reward $r_t \sim \mathcal{N}(m_{i_t}, \sigma_s)$. The decision-maker's objective is to choose a sequence of arms $\{i_t\}_{t \in \{1,\ldots,T\}}$ that maximizes cumulative expected reward $\sum_{t=1}^{T} m_{i_t}$, where T is the horizon length of the sequential allocation process.

For a multi-armed bandit problem, the performance of a decision-making policy is characterized in terms of the expected *regret* at time t defined by $R_t = m_{i^*} - m_{i_t}$, where $m_{i^*} = \max\{m_i \mid i \in \{1, \ldots, N\}\}$. The objective of the decision-maker can be equivalently defined to be to minimize the cumulative expected regret defined by $\sum_{t=1}^{T} R_t = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i^T]$, where n_i^T is the cumulative number of times option i has been chosen up to time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm i instead of arm i^*.

2.1 Bound on Optimal Performance

Lai and Robbins [9] showed that any asymptotically efficient algorithm for the multi-armed bandit problem must chose each suboptimal arm i at a rate that is at least logarithmic in time, i.e.,

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

where $o(1) \to 0$ as $T \to \infty$ and $D(\cdot||\cdot) \mapsto \mathbb{R}_{\geq 0} \cup \{+\infty\}$, defined by

$$D(p_i||p_{i^*}) = \int p_i(r) \log \frac{p_i(r)}{p_{i^*}(r)} \mathrm{d}r,$$

is the Kullback-Leibler divergence between the reward density p_i of the suboptimal option i and the reward density p_{i^*} of the optimal arm. For the Gaussian reward structure considered in this paper, the Kullback-Leibler divergence is equal

to $D(p_i||p_{i^*}) = \Delta_i^2/2\sigma_s^2$, and consequently, $\mathbb{E}[n_{iT}] \geq (2\sigma_s^2/\Delta_i^2 + o(1)) \log T$. This leads to a lower bound on the cumulative regret given by

$$\sum_{t=1}^{T} R_t \geq \sum_{i=1}^{N} \Big(\frac{2\sigma_s^2}{\Delta_i} + o(1) \Big) \log T.$$

2.2 Upper Credible Limit Algorithm for Gaussian Bandits

Let the prior on the mean reward at arm i be a Gaussian random variable with mean μ_{i0} and variance σ_0^2. Let the empirical mean of the rewards from arm i until time t be $\bar{m}_i^t$. Conditioned on the number of visits n_i^t to arm i and the empirical mean $\bar{m}_i^t$, the posterior distribution of the mean reward (M_i) at arm i at time t is a Gaussian random variable with mean and variance

$$\mu_i^t := \mathbb{E}[M_i | n_i^t, \bar{m}_i^t] = \frac{\delta^2 \mu_i^0 + n_i^t \bar{m}_i^t}{\delta^2 + n_i^t}, \text{ and } \sigma_i^{t\,2} := \text{Var}[M_i | n_i^t, \bar{m}_i^t] = \frac{\sigma_s^2}{\delta^2 + n_i^t},$$

respectively, where $\delta^2 = \sigma_s^2/\sigma_0^2$. Moreover,

$$\mathbb{E}[\mu_i^t | n_i^t] = \frac{\delta^2 \mu_i^0 + n_i^t m_i}{\delta^2 + n_i^t} \text{ and } \text{Var}[\mu_i^t | n_i^t] = \frac{n_i^t \sigma_s^2}{(\delta^2 + n_i^t)^2}.$$

For the Gaussian multi-armed bandit problem, we develop the upper credible limit (UCL) algorithm. The UCL algorithm at each (discrete) time t first computes the $(1 - 1/Kt)$-upper credible limit Q_i^t associated with each arm $i \in \{1, \ldots, N\}$ defined by

$$Q_i^t := \mu_i^t + \frac{\sigma_s}{\delta^2 + n_i^t} \Phi^{-1}\Big(1 - \frac{1}{Kt}\Big),$$

where $K > 0$ is a constant and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for the standard normal random variable. The UCL algorithm then selects an arm $i_t := \arg\max\{Q_i^t \mid i \in \{1, \ldots, N\}\}$. We show that for the non-informative prior, i.e., $\delta^2 \to 0^+$, the UCL algorithm achieves a logarithmic regret for a multi-armed bandit with Gaussian rewards. In particular, the regret satisfies the following uniform upper bound:

$$\sum_{t=1}^{T} R_t^{\text{UCL}} \leq \sum_{i=1}^{N} \Delta_i \left(\Big(\frac{8\sigma_s^2}{\Delta_i^2} + 2 \Big) \log T + \frac{4\sigma_s^2}{\Delta_i^2} (-\log\log T + 1 - \log 2) + 2 \right),$$

where R_t^{UCL} is the regret of UCL algorithm at time t. Moreover, for a good prior, i.e., a prior with small value of $|\mu_{i0} - m_i|/\sigma_0$, a similar performance can be achieved. In contrast, for a bad prior, i.e., a prior with large value of $|\mu_{i0} - m_i|/\sigma_0$, the expected cumulative regret may have super-logarithmic growth. The parameter K is chosen as $\sqrt{2\pi e}(\log T)^c$, where $c \geq 0$. For the non-informative prior $c = 0$ achieves a logarithmic regret, while for other good priors, the parameter c needs to be tuned to achieve a logarithmic regret. In the case of a bad prior, regret may have super-logarithmic growth independent of c.

2.3 Incorporating decision noise in the UCL algorithm

The UCL algorithm described in the previous section deterministically selects an arm with maximum upper credible limit. Human decision-making is inherently noisy and the soft-max action selection rule has been used in the literature to capture the noise in human decision-making. Accordingly, we consider a stochastic arm selection policy where at time t, arm i is selected with probability proportional to $\exp(\eta_t Q_i^t)$, where η_t is a cooling schedule of the form $\eta_t = \gamma \log t$, $\gamma > 0$. We show that there exists a feedback law for the cooling schedule parameter γ such that the stochastic arm selection policy achieves a logarithmic regret for good priors.

3 Comparison with human decision-making in multi-armed bandits

Human decision-making in multi-armed bandit problems has been studied in the cognitive psychology literature; see [5, 2, 1, 12, 15], and references therein. The salient features of the human decision-making are: (i) the decision at time t is based on a linear combination of the estimate of the mean reward of each arm and an ambiguity bonus that captures the value of the information from that arm [14], (ii) the decision-making is inherently noisy [14, 5, 1, 12, 15], (iii) the exploration-exploitation trade-off is sensitive to the time-horizon T for the bandit task [5, 14], (iv) the familiarity with the environment and the structure of the environment plays a critical role [5, 12, 2].

The value function Q_i^t of the UCL algorithm is comprised of two components: the first component is the estimate of the mean reward from arm i, while the second component is half the width of the minimum $(1 - 1/Kt)$-credible set. Since

the width of the minimum $(1 - 1/Kt)$-credible set is a measure of the ambiguity in the estimate of the mean reward, the value function Q_t^i well captures feature (i) of human decision-making. Furthermore, the expression for Q_i^t highlights that for efficient performance the desired credibility of the minimum credible set should increase with time.

Feature (ii) of human decision-making is the decision-making noise that is well captured by the stochastic arm selection policy. Our analysis of the soft-max arm selection policy highlights the role of decision-making noise in the performance. In particular, there exists a cooling schedule that achieves efficient performance for the stochastic arm selection policy, while for an arbitrary/non-diminishing decision noise the performance may be bad (i.e., the regret may be super-logarithmic).

Feature (iii) of human decision-making pertains to sensitivity to the time horizon and is captured by the model by the parameter $K = \sqrt{2\pi e}(\log T)^c$. In particular, the desired credibility $(1 - 1/Kt)$ increases with time horizon T, which results in higher value of Q_i^t for uncertain options and results in higher exploration.

Feature (iv) of human decision-making is captured through the priors in the UCL algorithm. In particular, familiarity with the environment results in a good prior and hence, efficient performance. We also show that for very good priors (that correspond to experts) even sub-logarithmic regret can be achieved. The structure of the environment can be modeled by assuming that the mean rewards of arms are sampled from a correlated Gaussian process, and the structural learning in human decision-making is captured by assuming a correlated prior. The UCL algorithm immediately extends to the case of correlated priors by replacing the univariate estimation procedure for each arm by a single multivariate estimation procedure for all arms and computing the upper credible limits using marginal distributions. In particular, for spatially distributed multi-armed bandits, a normal prior with mean $\mu_0 \mathbf{1}_N$ and covariance $\Sigma_{ij} = \sigma_0^2 \exp(-|x_i - x_j|/\lambda)$, where x_i is the location of arm i and $\lambda \geq 0$ is the correlation length scale parameter, well captures a variety of possible correlation structures (true structures as well as structures believed by the human). In particular, as λ varies from 0 to $+\infty$, the correlation structure varies from independent to completely correlated.

4 Conclusions

We considered the multi-armed bandit problem with Gaussian rewards and developed UCL algorithm that well models human decision-making in such tasks. We showed that the UCL algorithm achieves logarithmic regret for good priors and appropriate cooling schedules, while bad priors and inefficient cooling schedules may result in super-logarithmic regret. The proposed model is primarily parametrized by four parameters, namely, the cooling schedule parameter γ, prior estimates of mean μ_0, prior variance σ_0^2 and correlation parameter λ. Variations in these parameters capture a variety of empirically observed behaviors in human decision-making.

References

[1] D. Acuña and P. Schrater. Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, volume 100, pages 200–300, Washington, DC, USA, July 2008.

[2] D. E. Acuña and P. Schrater. Structure learning in human sequential decision-making. *PLoS computational biology*, 6(12):e1001003, 2010.

[3] R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[5] J. D. Cohen, S. M. McClure, and A. J. Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

[6] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

[7] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[8] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Int Conf on Artificial Intelligence and Statistics*, pages 592–600, La Palma, Canary Islands, Spain, April 2012.

[9] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[10] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

[11] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[12] M. Steyvers, M. D. Lee, and E. Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, 2009.

[13] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[14] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Why the grass is greener on the other side: Behavioral evidence for an ambiguity bonus in human exploratory decision-making. In *Neuroscience 2011 Abstracts*, Washington, DC, November 2011.

[15] S. Zhang and A. J. Yu. Cheap but clever: Human active learning in a bandit setting. In *Proceedings of the Cognitive Science Society Conference*, Berlin, Germany, August 2013. to appear.