Bandits Problems

Extrait de : Adrien Goëffon, Frédéric Lardeux, Frédéric Saubion

Simulating non-stationary operators in search algorithms.
Applied Soft Computing 38: 257-268, Elsevier, (2016)

1 Multi-Armed Bandits

The initial stochastic multi-armed bandit (MAB) [14, 3, 15] is formulated as follows. Given several possible actions — usually called arms according to the gambling machine analogy — that have different individual gains (or rewards), one has to select a sequence of actions that maximizes the total gain. Definition 1 proposes a more formal definition of this general problem.

Definition 1 (Stochastic MAB) *Let us consider n independent arms. For each arm $i \in \{1, \dots, n\}$, we have:*

- *a set of possible states S_i;*
- *a set of probabilities $Prob_i = \{\sigma^i_{j \to k} | j, k \in S_i\}$ such that $\sigma^i_{j \to k}$ is the probability of being in state k if the arm i is played*[1] *from state j;*
- *a set of gains $G_i = \{g^i_j | j \in S_i\}$ where g^i_j is the gain obtained when arm i is played from state j.*

Given a stochastic MAB, the problem is to find a policy that maximizes over a finite[2] horizon T, $\sum_{t=0}^{T} g_t \gamma^t$, where g_t is the expected gain of the policy at time t and $\gamma \in [0, 1]$ is a discount factor.

Four features can be identified to characterize a MAB problem [13]:

1. only one arm is played at each time;
2. states of unplayed arms do not change;
3. arms are independent;
4. arms that are not played do not contribute any gain.

[1] We use the verb *play* according *gain* to the gambling analogy.

[2] Note that we restrict the problem to finite horizon MAB. The most general problem is often presented over infinite horizon.

Many variants of the initial stochastic MAB have been studied in the literature. In this paper we focus on the restless MAB, first introduced in [18]. In this formulation, the gains of the arms change over time, while they are supposed to be fixed — but of course unknown — in the initial stochastic MAB formulation. In fact, restless bandits may be defined as in Definition 1 except that, when an arm is not played, its state may change, which corresponds to a relaxation of Feature 2. Hence, restless MABs involve two kinds of probabilities in $Prob_i$, namely $\sigma^i_{j\rightarrow k}$, which represents the probability of being in state k if the arm i is played, and $\tilde{\sigma}^i_{j\rightarrow k}$, that is the probability of being in state k if the arm i is not played.

OS policy problems can be stated as MAB problems, where arms represent operators. In order to simulate the behavior of OS policies, one may define specific scenarios within the MAB formalism that specify gain and probabilities of gains of operators.

2 Operator Selection Policies

In this section we first explain how operator selection in operator based algorithms can be directly related to the choice of the most suitable sequence of actions in the context of multi-armed bandit problems. We review then different possible selection policies that can be used to achieve an optimal schedule of the operators.

2.1 Operator Selection in Operator Based Algorithms

Let us consider an adaptive OBA $A = (Init, \Omega, \theta, \pi, K)$. Here, we are not interested in the initialization function $Init$. Let $\Omega = \{o_1, \dots, o_n\}$ be the set of n operators. We have to define the control policy (π, K) which selects an operator at each iteration of the algorithm in order to build a run $(\overline{s}, \overline{o})$. We review here different policies and we distinguish between policies based on probabilities of application of the operators and policies based on upper confidence bounds.

The gain of an operator is generally specific to the problem, since it uses the notion of performance of a run. In order to have a more general approach, a general notion of utility[3], which reflects the successive gains obtained by the operators, can be introduced.

Considering a run $(\overline{s}, \overline{o})$, such that $\overline{s} = s^{(0)}, \dots, s^{(n)}$ and $\overline{o} = o^{(1)}, \dots, o^{(n)}$, an utility $u_i^{(t)}$ is associated to each operator $i \in \{1..n\}$ for any iteration $t \in \{1..n\}$. This utility has to be re-evaluated at each time, classically using a formula $u_i^{(t)} = (1-\alpha)u_i^{(t-1)} + \alpha . g(o_i, s^{(0)}, \dots, s^{(t-1)})$, with $u_i^{(0)} = 0$. This utility uses the gain associated to the application of operator i (which corresponds thus to the immediate utility) and α which is a coefficient that controls the balance between past and immediate utilities, as in classic reinforcement learning techniques [16]. If an operator is not selected at iteration t, its gain is 0 for this iteration.

[3] Note that we use the term utility here, which should be clearly related to the notion of action value in reinforcement learning [16].

2.1.1 Policies based on probabilities of application

In this context, given the set of operators $\Omega = \{o_1, \ldots, o_n\}$, we use the parameter vector θ to associate a probability of selecting the operator, $\theta = (\sigma_1, \ldots, \sigma_n)$ such that $\sum_{i=1}^n \sigma_i = 1$. The OS policy π is then a roulette selection wheel that selects each operator o_i according to its probability of selection σ_i. Different operator selection policies have been proposed in the literature [12, 9]; we review here some of the most used of them.

- **Fixed Roulette Wheel**

  A first possibility consists in keeping θ fixed during the run, i.e. $\forall t \in \mathbb{N}, K(\theta, t) = \theta$. Note that these values can be determined by an automated tuning process [11, 5].

- **Adaptive Roulette Wheel**

  Contrary to a static tuning of the operator application rates, adaptive operator selection consists in selecting the next operator to apply at iteration $t+1$ by adapting the selection probability during the search. In this case, we have $\theta^{(t)} = (\sigma_1^{(t)}, \ldots, \sigma_n^{(t)})$. The control function $K : \Theta \times \mathbb{N} \rightarrow \Theta$ is defined as $K(\theta^{(t)}, t+1) = \theta^{(t+1)}$. Defining K consists in defining the probabilities $\sigma_i^{(t+1)}$ with regards to the evolution of the operator's utilities.

  A classic mechanism is the probability matching selection rule:

  $$\sigma_i^{(t+1)} = p_{\min} + (1 - n.p_{\min}) \frac{u_i^{(t+1)}}{\sum_{k=1}^n u_k^{(t+1)}} \tag{1}$$

  where a non-negative value $p_{\min}$ insures a non zero selection probability for all operators. Note that, in order to insure a coherent behavior, $p_{\min}$ should be in the interval $[0, \frac{1}{n}]$.

- **Adaptive Pursuit**

  An alternative proportional selection rule has been proposed in [17], called adaptive pursuit (AP), that distinguishes the best current operator from the others:

  $$\begin{cases} \sigma_{i^*}^{(t+1)} = \sigma_{i^*}^{(t)} + \beta(p_{\max} - \sigma_{i^*}^{(t)}) \\ \sigma_i^{(t+1)} = \sigma_i^{(t)} + \beta(p_{\min} - \sigma_i^{(t)}) \end{cases} \tag{2}$$

  where $i^* \in \underset{i \in \{1,\ldots,n\}}{\text{argmax}} \; u_i^{(t+1)}$, $p_{\max} = 1 - (n-1)p_{\min}$ and β is a parameter to adjust balance of this winner-take-all strategy.

2.1.2 Policies based on upper confidence bounds

Optimal strategies have been initially proposed by [6] and [8] for the multi-armed bandit problem. Later, [1] proposed to use this problem to manage the compromise between exploration and exploitation in optimization algorithms. The following policies consists in computing an upper confidence bound of the expected gain and to select thus the most promising arm.

- **UCB (upper confidence bound)**

  The UCB1 criterion [2] is defined as:

$$\forall o_i \in \Omega, UCB1(o_i, t) = u_i^{(t)} + \sqrt{\frac{2\log(\sum_{1\leq k\leq n} nb_k^{(t)})}{nb_i^{(t)}}} \tag{3}$$

  where $nb_i^{(t)}$ denotes the number of times operator o_i has been applied. Note that this formula is defined for gains that should be normalized between 0 and 1. The left term of the formula uses the successive utilities that are obtained by the arms in order to focus of the best arm, while the right term aims at providing the opportunity to be selected for less used arms. This formula attempts thus to achieve a compromise between exploitation and exploration.

  Therefore, we may define the control policy (π, K) for a given iteration t as:

$$\begin{cases} \pi(\theta^{(t)}, t) = \underset{i\in\{1,\ldots,n\}}{\text{argmax}}\ \theta_i^{(t)} \\ K(\theta^{(t)}, t) = \theta^{(t+1)} = (UCB1(o_1, t+1), \ldots, UCB1(o_n, t+1)) \end{cases} \tag{4}$$

  Here no parameter is required. Note that UCB has originally be designed for fixed gain distributions. Since the gain of the operators is likely to change along the search, UCB has been extended to dynamic multi-armed bandit has to be considered.

- **DMAB (Dynamic MAB algorithm based on UCB)**

  UCB has been revisited in [4]. A standard test — known as Page Hinkley [10] — for the change hypothesis is used. We may add a parameter in θ which indicates if the process has to be restarted. In this case the control function K use the Page Hinkley test to detect statistical changes in the successive utilities of the operators and may re-initialize the values of the operators utilities. Moreover, it can be useful to add a scaling factor to the right term of the UCB1 formula in order to take into account the value range for utilities. The test is parametrized by γ that controls its sensitivity and δ that manages its robustness. We refer the reader to [7] for more details.

References

[1] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[3] R. N. Bradt, S. M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics*, 27(4):1060–1074, 1956.

[4] L. Da Costa, Á. Fialho, M. Schoenauer, and M. Sebag. Adaptive operator selection with dynamic multi-armed bandits. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'08)*, pages 913–920. ACM, 2008.

[5] A.E. Eiben and S.K. Smit. *Autonomous Search*, chapter Evolutionary Algorithm Parameters and Methods to Tune them, pages 25–38. Springer, 2012.

[6] D. Feldman. Contributions to the "two-armed bandit" problem. *The Annals of Mathematical Statistics*, 33(3):847–856, 1962.

[7] Álvaro Fialho, Luís Da Costa, Marc Schoenauer, and Michèle Sebag. Analyzing bandit-based adaptive operator selection mechanisms. *Ann. Math. Artif. Intell.*, 60(1-2):25–64, 2010.

[8] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[9] Youssef Hamadi, Eric Monfroy, and Frédéric Saubion. *Autonomous search.* Springer-Verlag, 2012.

[10] David V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.

[11] Holger H. Hoos. *Autonomous Search*, chapter Automated Algorithm Configuration and Parameter Tuning, pages 37–71. Springer Verlag, 2012.

[12] F. Lobo, C. Lima, and Z. Michalewicz, editors. *Parameter Setting in Evolutionary Algorithms*, volume 54 of *Studies in Computational Intelligence.* Springer, 2007.

[13] Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In III Hero, AlfredO., DavidA. Castañón, Douglas Cochran, and Keith Kastella, editors, *Foundations and Applications of Sensor Management*, pages 121–151. Springer US, 2008.

[14] H. Robbins. Some aspects of the sequential desing of experiments. *Bulletin Amrican Mathematical Society*, (55):527–535, 1952.

[15] L. Rodman. On the many-armed bandit problem. *The Annals of Probability*, 6(3):491–498, 1978.

[16] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.

[17] D. Thierens. An adaptive pursuit strategy for allocating operator probabilities. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'05)*, pages 1539–1546. ACM, 2005.

[18] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.