ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

The Exploration vs Exploitation Trade-Off in Bandit Problems: An Empirical Study

Saba Yahyaa and Bernard Manderick

Vrije Universiteit Brussel - Computational Modeling Lab
Pleinlaan 2, B-1050 Brussels - Belgium

Abstract. We compare well-known action selection policies used in reinforcement learning like ϵ-greedy and *softmax* with lesser known ones like the Gittins index and the knowledge gradient on bandit problems. The latter two are in comparison very performant. Moreover the knowledge gradient can be generalized to other than bandit problems.

1 Introduction

In the multi-armed bandit problem (or bandit problem for short), introduced by Robbins [10], the agent has to decide at each time step which one of K different arms to choose in order to maximize its total expected reward. For each arm, the rewards are generated according to a given family of distributions with unknown parameters, e.g. the family of normal distributions.

The bandit problem has received a lot of attention since it reflects the essence of the *trade-off* between *exploration*, i.e. to collect enough information about all arms in order to learn which one is best, and *exploitation*, i.e. to use that information to avoid the underperforming arms. If the parameters of the reward distributions are known, then it is optimal to always select the arm with the highest mean. However, since these distributions are unknown, the agent has to explore the different arms hoping not to spend too much time exploring arms which are not the best ones. Often simple heuristics are used as *action selection policy*, e.g. the ϵ-greedy policy selects the arm believed to be best most of the time while every now and then another arm is selected to collect information about it [12].

In this paper we evaluate empirically a number of action selection policies including the Gittins index policy and the knowledge gradient policy on a test set of bandit problems. Although trading-off exploration and exploitation is very important for reinforcement learning in general and the bandit problem reflects the essence of this trade-off, only few empirical studies have been published, e.g. [13], and we are aware about only one systematic empirical evaluation [8].

The rest of the paper is organized as follows. In Section 2, we present the bandit problem and in Section 3, the action selection policies are introduced. We present our results and conclude in Section 4.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 The Bandit Problem

In the bandit problem, the environment remains forever in the single *physical* state s regardless which one of the K actions[1] $i = 1, \cdots, K$ are taken. The rewards of the actions i are drawn from normal distributions with different means μ_i but common standard deviation σ. However, these parameters are *unknown* to the agent. In this paper we focus on the *infinite horizon* with geometrical discount factor γ, i.e. the agent has to maximize the total expected reward $\mathbb{E}(\sum_{n=0}^{\infty} \gamma^n r_n)$ where r_n is the reward obtained at time step n since the Gittins index theorem, a strong theoretical result, holds for this case.

Already, Bellman [2] realized that the action selection problem can be formulated as a dynamic program. Therefore, the agent maintains K *knowledge states*[2], one for each arm reflecting the agent's knowledge about that arm. At each time step n, the following information is available: 1) the number of times $n_i(n)$ each arm i has been tried[3], and 2) the estimated mean $\hat{\mu}_i(n)$ and variance $\hat{\sigma}_i(n)$ of the corresponding reward distribution based on these n_i trails. Assume that at time step n, arm i is selected and reward r_{n+1} is received. The corresponding knowledge state $(n_i, \hat{\mu}_i, \hat{\sigma}_i^2)$ is updated as follows while the other states remain unchanged:

$$n_{i+1} = n_i + 1 \tag{1}$$

$$\hat{\mu}_{i+1} = (1 - \frac{1}{n_i})\hat{\mu}_i + \frac{1}{n_i} r_{n+1} \tag{2}$$

$$\hat{\sigma}_{i+1}^2 = \frac{n_i - 2}{n_i - 1}\hat{\sigma}_i^2 + \frac{1}{n_i}(r_{n+1} - \hat{\mu}_i)^2 \tag{3}$$

Another important quantity is the variance $\hat{\bar{\sigma}}_i^2$ of the mean $\hat{\mu}_i$, i.e. $\hat{\bar{\sigma}}_i^2 = \hat{\sigma}_i^2 / n_i$. If $\hat{\bar{\sigma}}_i^2$ is low, our confidence in our estimate $\hat{\mu}_i$ is high.

But now, the agent has to maintain K knowledge states and when the number of arms K is large, one faces the 'curse of dimensionality' [2]. However, in case of the bandit problem it is possible to compute for each arm its Gittins index and this independently from the other ones. And, the optimal policy is to choose the arm with the highest index. This way, a K-dimensional problem is reduced to K 1-dimensional problems and there is no curse as proven in Gittins and Jones [5].

Unfortunately, this result is only valid for bandit problems with certain properties and cannot be generalized easily. Recently, a heuristic inspired by the Gittins index policy that can be generalized was introduced: the knowledge gradient [6, 4].

[1] In order to simplify mathematical notation, actions are represented by their index, i.e. $1, \cdots, K$ instead of $a_1, \cdots, a_K$.

[2] Bellman used information state instead of knowledge state.

[3] In order not to overload the notation we omit the time step n when it does not cause confusion.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

3 Action Selection Policies

In our study, we compared 8 action selection policies. The first 5 are often used in reinforcement learning, i.e. ϵ-greedy and softmax exploration, pursuit, reinforcement comparison, and interval esitmation (IE). For more details about these policies, we refer to [7] for interval estimation and to [12] for the rest. We also included 3 lesser known ones described next: upper confidence bound, Gittins index, and knowledge gradient policies.

3.1 Upper Confidence Bound Policy (UCB)

The *upper confidence bound* (or *UCB*) policy was proposed in [1]. Actually, it is a family of policies of which we consider two members: *UCB1* and *UCB1-Tuned*.

The *UCB1* policy first plays each arm once. From then on, at each time step n it selects the arm j that maximizes the function $\hat{\mu}_i + \sqrt{2 \ln n / n_i}$. The authors also introduced *UCB1-Tuned* that apart from the estimated means $\hat{\mu}_i$, also takes the estimated variances $\hat{\bar{\sigma}}_i^2$ of these means into account. Here, the arm is selected according to

$$j = \underset{i=1,\cdots,K}{\arg\max} \hat{\mu}_i + \sqrt{\frac{\ln n}{n_i} min\{\frac{1}{4}, V_i(n_i)\}} \text{ where } V_i(n) = \hat{\bar{\sigma}}_i^2 + \sqrt{\frac{2 \ln n}{n_i}}$$

In [1] it is shown that *UCB1* is optimal in terms of regret. For *UCB1-Tuned* no such theoretical guarantees exist although according to the authors it performs better than *UCB1* in practice. So, we also included it in our comparison.

3.2 Gittins Index Policy

The famous *Gittins index* theorem [5] states that for the multi-armed bandit problem with geometric discount and independent arms, it is *optimal* at each time step to select the arm with the highest index ν_i which depends on the number of times n_i that arm has been selected.

Unfortunately, the Gittins index is hard to compute but in case of rewards drawn from the standard[4] normal distribution $N(0, 1)$, tables exist and for any other normal distribution $N(\mu, \sigma)$, the corresponding Gittins index can be computed from these tables [5, 9]. Also, good approximations for the upper and lower bound for the Gittins index exist, e.g. Brezzi and Lai [3] propose for discrete time problems the approximation $\Gamma(n) = \frac{P^G}{\sqrt{n}}$ where P^G is a parameter to be tuned. For more information on the theory and different proofs of the Gittins index theorem, we refer to [5].

3.3 The Knowledge Gradient Policy (KG)

The *knowledge gradient policy* (or *KG*) was first introduced in Gupta and Miescke [6] and further analyzed by Frazier and Powell for online learning [4, 9],

[4] The standard normal distribution $N(0, 1)$ has mean $\mu = 0$ and variance $\sigma = 1$.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

who also proposed the name knowledge gradient. KG chooses the action i with the largest value of $V^{KG}(i)$ and it prefers those actions about which comparatively little is known. These actions are the ones whose distributions around the estimate mean $\hat{\mu}_i$ have larger standard deviations $\hat{\tilde{\sigma}}_i$. Thus, KG prefers an action i over its alternatives if its confidence in the estimate mean $\hat{\mu}_i$ is low. For the multi-armed bandit problem, which is an online problem, this policy selects the next action according to

$$J_{KG}(n) = \arg\max_i \hat{\mu}_i + (n - n_i) V^{KG}(i) \text{ where } V^{KG}(i) = \hat{\tilde{\sigma}}_i \, f(-|\frac{\hat{\mu}_i - max_{j \neq i} \hat{\mu}_j}{\hat{\tilde{\sigma}}_i}|)$$

Here, $f(\zeta) = \zeta \Phi(\zeta) + \phi(\zeta)$ where $\Phi(\zeta)$ and $\phi(\zeta)$ are the cumulative distribution and the density of the standard normal density, respectively.

4 Empirical Comparison

The experimental setup is as follows. The number of arms was varied as follows: $K = 2, 5, 10, 50$. The rewards of the K arms are all distributed normally with the mean rewards μ_i generated according to a uniform distribution over the closed interval $[0, 1]$ but each arm has the same variance σ. The means were sorted from high to low so that the first arm was always the best while the last one the worst, i.e. $\mu_1 > \mu_2 > \cdots > \mu_K$. We experimented with standard deviations[5] $\sigma = 0.01, 0.1, 1.0$. The smaller σ, the easier the problem. We also considered the case where arms have different but fixed σ_i. The σ_i are generated from a uniform distribution over the closed interval $[0, 1]$. Many of the action selection policies described in Section 3 have tunable parameters which are optimized using cross-validation and we compare policies with their optimal parameter values. As in [8], we used the performance measures 1) *Total regret accumulated over the experiment*, 2) *Regret as a function of time*, and, 3) *Percentage of plays in which the optimal arm is pulled*, but we only show results for the last one.

4.1 Comparison of Action Selection Policies

Our results concerning the 5 first action selection policies described in Section 3 are consistent with the ones reported in [8]. Most importantly and quite surprisingly the simplest policies, ϵ-greedy and *softmax*, are almost always better than the rest with *softmax* slightly better than ϵ-greedy. Only for a small number of arms ($K = 2, 5$) and high standard deviation, i.e. $\sigma = 1$, is *UCB1-Tuned* doing better. Since this policy takes also the variance into account, cf. Subsection 3.1, this should not come as a surprise. The simple pursuit policy is the worst.

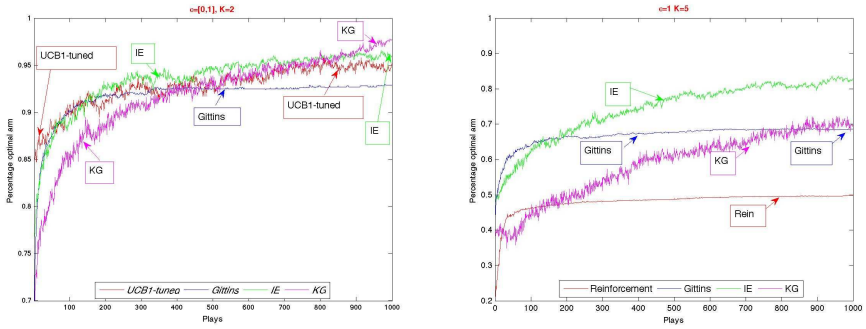
The second conclusion is that the performance of a policy very much depends on the parameter values used. So, these values should be optimized first before one compares them with other policies.

Next, we compared the performance of the Gittins index, interval estimation, and knowledge gradient policies with the best action selection policy from [8] for

[5]The standard deviation squared equals the variance.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

each combination of arms K and standard deviations σ. The results are shown in Figure 1 for the *Percentage of optimal action* performance measure. For the two other ones, the results are similar but they are not shown here because of lack of space. Before we applied these new policies we tuned optimally the involved parameter, i.e. P^G for the Gittins index and Z_α for the interval estimation policy. An advantage of the knowledge gradient policy is that no parameters have to be tuned.



(a) K=2, $\sigma \in I$ (b) K=5, $\sigma = 1$

(c) K=10, $\sigma \in I$ (d) K=50, $\sigma = 1$

Fig. 1: Comparison of action selection policies on a test set of bandit problems, cf. text for explanation.

In Figure 1, we show the comparison of the Gittins index, IE, and KG policies with the best one from [8]. In subfigure (a), K=2, σ in the interval $I = [0, 1]$, the best policy is KG. In subfigure (b), K=5, $\sigma = 1$, the best policy is IE. In the subfigure (c), K=10, $\sigma \in I$, the best policy is KG, and in the subfigure (d), K=50, $\sigma = 1$, the best policy is KG. We have done the comparison on 2, 5, 10, and 50-armed bandit problems for a range of standard deviations and the obtained results are similar. The most important observations of this empirical comparison can be summarized as follows:

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

For $K = 2$ as long as the variances are fixed the 3 new policies perform more or less the same and outcompete the best one in the study of Kuleshov and Precup [8] but when the arms have different variances *Gittins index* is much worse than the others. This is quite surprising and so far we do not have an explanation for that. For $K = 5, 10$, the results are more or less the same. However, for many arms and/or high variance of the rewards and arms with different variances, KG is outperforming the rest.

The main conclusion of this study is quiet simple. The *knowledge gradient* policy (KG) is in all circumstances always at least as good as the most competitive new policies. In case of many arms or high variance, KG is the clear winner. Moreover, its computational complexity is low and no parameters have to be tuned.

References

[1] P. Auer, N. Cesa-Bianchi and P. Fischer, Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, 47:235-256, Kluwer Academic Publishers, May, 2002.

[2] R. Bellman, A Problem in the Sequential Design of Experiments, *Sankhya: The Indian Journal of Statistics*, 16:221-229, Springer, 1956.

[3] M. Brezzi and T. L. Lai, Optimal learning and experimentation in bandit problems, *Economic Dynamics and Control*, 27(1):87-108, Elsevier November 2002.

[4] P. Frazier, W. B. Powell and S. Dayanik, A Knowledge-Gradient Policy for Sequential Information Collection, *SIAM J. Control and Optimization*, 47:2410-2439, 2008.

[5] J. C. Gittins, K. Glazebrook and R. Weber. *Multi-Armed Bandit Allocation Indices*, J. Wiley And Sons, Series Wiley-Interscience Series In Systems And Optimization, second Edition, New York, 2011.

[6] S. Gupta and K. Miescke, Bayesian look ahead one stage sampling allocations for selecting the largest normal mean, *Statistical Papers*, 35:169-177, Springer, 1994.

[7] L. P. Kaelbling. *Learning in Embedded Systems*, The MIT Press, 1993.

[8] V. Kuleshov and D. Precup, Algorithms For Multi-Armed Bandit Problems, *Machine Learning Research*, October 2010.

[9] W. B. Powell. *Approximate Dynamic Programming: Solving The Curses Of Dimensionality*, J. Wiley And Sons, New York, 2007.

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

[11] I. O. Ryzhov, W. Powell and P. I. Frazier, The knowledge gradient algorithm for a general class of online learning problems, *Operation Research*, 0(0):1-33, 2008.

[12] R. S. Sutton and A.G. Barto, *Reinforcement learning: An introduction*, The MIT Press, 1998.

[13] J. Vermorel and M. Mohri, Multi-armed Bandit Algorithms and Empirical Evaluation. In J. Gama, R. Camacho, P. Brazdil, A. Jorge and L. Torgo, editors, proceedings of the 16^{th} *European Conference on Machine Learning* (ECML 2005), Lecture Notes in Computer Science 3720, pages 437-448, Springer, 2005.