Operations Research Letters 43 (2015) 143–147

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Refined knowledge-gradient policy for learning probabilities

Bogumił Kamiński *

Warsaw School of Economics, Al. Niepodległości 162, 02-554 Warszawa, Poland

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form
31 December 2014
Accepted 1 January 2015
Available online 8 January 2015

Keywords:
Ranking and selection
Knowledge gradient
Learning probabilities
Bayesian inference

ABSTRACT

In a ranking and selection problem with independent Bernoulli populations we study the knowledge-gradient policy. We show that this policy is inefficient and propose three alternative rules that have superior performance. In particular we show that one of the considered policies – called KG(∗) in the literature – is numerically difficult and expensive to evaluate and it is possible to derive a much simpler policy that gives comparable results.

1. Introduction

The knowledge-gradient (KG) policy for ranking and selection problems was first proposed by Gupta and Miescke [7] and subsequently developed by Frazier et al. [4]. The objective is to choose from $N \geq 2$ alternatives the one that has the highest expected performance. It is assumed that we can sequentially sample the alternatives to learn their unknown performance but each measurement is noisy. The key development assumption behind the KG policy in relation to alternative policies for ranking and selection is its simplicity [4].

Ranking and selection problems are encountered in numerous practical situations, see for example [2] for a review. In the standard approach we assume that the number of possible alternatives N is small and that they are independent, i.e. observing one alternative does not give us information about other alternatives (notably, however, Frazier et al. [5] propose an algorithm that allows to take into account dependencies between alternatives). The main stream of research on the KG policy [4,2,3] assumes that the performance measure is a real number and measurement noise is normally distributed. Recently Powell and Ryzhov [8] extend it to other continuous (exponential, uniform) and discrete (Poisson, Bernoulli) distributions of measurements. In particular they highlight that the Bernoulli distribution case – implying that we want to learn the probability of success – was a setting motivating the development of the optimal learning methods. However, it was earlier omitted in the literature on the KG policy.

Learning probabilities is an important case of ranking and selection problems occurring in physical experiments (e.g. the probability of treatment success in clinical trials) and computer simulations (e.g. testing the probability that simulation output satisfies some condition). In this paper we will investigate the properties of the KG policy for such problems.

The idea of the KG policy is to sequentially sample alternatives that promise the highest increase of the expected value of performance measure of the best option. This one-step-ahead policy is known to be sub-optimal in comparison to the exact solution of a multiple-stage decision making problem, except in certain special cases [4]. However, it is reported to perform efficiently in practical applications for normally distributed measurements and its additional value lies in simplicity [4].

Unfortunately, the formula for the KG policy for learning probabilities (formula (5.26) on p. 112 in [8]) suggests that it might not be as effective in this case. Powell and Ryzhov [8] note that "*This hints at possible limitations of the knowledge gradient approach when our observations are discrete (...)*". However, they do not investigate the properties of the KG policy in the Bernoulli case in detail.

In this work we provide an analysis of the properties of the KG policy for learning probabilities and show that it can be expected to perform poorly. Using these findings we propose several refinements of this policy that significantly improve the learning rate. Firstly we analyze how KG(∗) policy, proposed by Frazier and Powell [2] for standard normally distributed measurement noise, can be applied in case of learning probabilities. It is found that it offers a much better performance than the KG policy. However, we show the KG(∗) policy is numerically difficult and expensive to

* Tel.: +48 22 564 60 00.
E-mail address: bkamins@sgh.waw.pl.

http://dx.doi.org/10.1016/j.orl.2015.01.001
0167-6377/

evaluate. Therefore we propose alternative policies that have almost the same performance, but are much simpler to compute.

The paper is structured as follows. In Section 2 we investigate the properties of the KG policy for learning probabilities. Next, in Section 3 we propose improved policies. They are compared using simulation in Section 4.

The computations presented in the text are performed using Java, GNU R and Python. Source codes and data sets allowing to replicate the analyses can be downloaded at http://bogumilkaminski.pl/pub/bkg.zip.

2. Properties of the knowledge gradient policy

Let $\{1, 2, \dots, N\}$ be the set of alternatives from which we want to select the best. Each alternative has some unknown probability of success that we maximize. These probabilities are not observable but we can run an experiment and obtain a sample from the set $\{0, 1\}$, where 1 is drawn with the unknown probability. For instance we might test N different drugs for which we do not know the probabilities of being effective. We can, however, administer the selected drug to a patient and observe if it was effective (1) or not (0).

Let us first describe the KG policy proposed by Powell and Ryzhov [8]. It is based on a Bayesian approach and assumes that we have a random variable $Y = (Y_1, Y_2, \dots, Y_N)$ representing our beliefs about the distribution of success probabilities of the alternatives. The sampling model for each alternative is beta-binomial, see for example [6]. We take it that the prior distributions of Y_i are independent and that Y_i has beta distribution with parameters α_i^0 and β_i^0.

Now let us assume that in step k we measure alternative i and observe $W_i^k \in \{0, 1\}$. Then, following Bayes' rule (see [8]), we know that the posterior distribution of Y_i is also beta with parameters:

$$\alpha_j^k = \alpha_j^{k-1} + W_i^k \quad \text{and} \quad \beta_j^k = \beta_j^{k-1} + 1 - W_i^k.$$

By the independence assumption, α_j^k and β_j^k for $j \neq i$ are not changed. Further we will denote $n_i^k = \alpha_i^k + \beta_i^k$, $p_i^k = \alpha_i^k / n_i^k$ and $c_i^k = \max_{j \neq i} p_j^k$. Notice that p_i^k is the expected value of probability of success of alternative i after k measurements and c_i^k is the expected value of the maximal expected probability of success for alternatives other than i. A Bayes-optimal decision if we stop experiments after k steps, is to choose an alternative for which $p_i^k \geq c_i^k$ (in general there might be more than one such an alternative).

The KG policy prescribes that after k steps we want to measure an alternative that maximizes the expected improvement of the performance of the alternative that would be selected after the measurement. The knowledge gradient v_i^k is defined as this expected improvement given that in step $k + 1$ alternative i is measured. The formula for v_i^k in beta-binomial case has the following form, see [8]:

$$\begin{cases} p_i^k \left(\dfrac{\alpha_i^k + 1}{n_i^k + 1} - c_i^k \right) & \text{if } p_i^k \leq c_i^k < \dfrac{\alpha_i^k + 1}{n_i^k + 1} \\ (1 - p_i^k) \left(c_i^k - \dfrac{\alpha_i^k}{n_i^k + 1} \right) & \text{if } \dfrac{\alpha_i^k}{n_i^k + 1} < c_i^k < p_i^k \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Powell and Ryzhov [8] notice that it is possible that v_i^k is equal to 0 for all i, in which case they recommend that the KG policy should choose an alternative at random (in this text we extend this recommendation to all situations when there is a tie between several alternatives with largest v_i^k). They give in Table 5.3 (p. 112) an extreme example of this case – where alternatives have strongly different numbers of measurements and probabilities of success –

Table 1
The mean and 95% CI of the probability that $\forall i : v_i^k = 0$ for $N = 5$ and $N = 10$.

k	$N = 5$	$N = 10$
25	72.44% (71.67%, 73.20%)	68.71% (67.94%, 69.46%)
50	81.24% (80.60%, 81.88%)	77.51% (76.94%, 78.08%)
100	87.60% (87.08%, 88.12%)	84.06% (83.59%, 84.52%)
200	92.10% (91.69%, 92.50%)	88.92% (88.53%, 89.30%)

Table 2
The mean and 95% CI of the probability that exactly one $v_i^k > 0$ given $\exists i : v_i^k > 0$ for $N = 5$ and $N = 10$.

k	$N = 5$	$N = 10$
25	40.41% (38.87%, 41.95%)	47.33% (46.12%, 48.54%)
50	46.63% (45.05%, 48.27%)	57.39% (56.23%, 58.59%)
100	50.51% (48.88%, 52.20%)	64.57% (63.37%, 65.77%)
200	52.69% (50.99%, 54.36%)	68.85% (67.64%, 70.04%)

and draw a conclusion that *"Intuitively, it can occur when we are already quite certain about the solution to the problem"*. In this paper we argue that in fact such a situation is common for the KG policy with binary outcome which results in its low efficiency. Therefore next we propose alternative policies that have better performance.

Let us first analyze when $v_i^k > 0$. From Eq. (1) we get the condition $\alpha_i^k/(n_i^k + 1) < c_i^k < (\alpha_i^k + 1)/(n_i^k + 1)$, which after rearrangement yields $\alpha_i^k < (n_i^k + 1)c_i^k < \alpha_i^k + 1$.

Now notice that the set A_i^k of possible values of α_i^k is a subset of the set $\{\alpha_i^0 + s : s \in \mathbb{N}\}$. Therefore there is at most one element in A_i^k that meets the above condition (in general this equation could even have no solution; for example when $\alpha_i^0 = 1$, $n_i^k = 2$ and $c_i^k = 0.1$). This observation suggests that the situation when $v_i^k > 0$ should have a low probability. This in turn means that we can expect that the KG policy reduces to a random search frequently. It is hard to provide a formal proof of this statement because values a_i^k, n_i^k and c_i^k are interdependent (they are generated conditional on the KG policy). Therefore we support this conjecture using a simulation.

We consider $N \in \{5, 10\}$ alternatives and $\alpha_i^0 = \beta_i^0 = 1$ (implying uniform a priori distributions of Y_i). We sample true success probabilities for each alternative independently and from a uniform distribution. We estimate the probability that $\forall i\colon v_i^k = 0$ during the process of optimization for the number of the KG algorithm steps equal to $k \in \{25, 50, 100, 200\}$ by 1000 replications of the simulation. The results are given in Table 1. We can notice that this probability is very high even for small k and is increasing with k. Additionally in Table 2 we show that if there exists a non-zero v_i^k there will be only one such positive value with a high probability that increases with k. The reason of such a situation is again that $\Pr(v_i^k > 0)$ becomes low very fast. We will use the latter observation when developing one of the alternatives to the KG policy.

Even considering the most extreme case when $N = 2$ and the true probabilities of success for alternatives 1 and 2 are both equal to 1, it can be checked by simulation that in 2/3 of cases $v_1^k = v_2^k = 0$ for large k.

In summary, the KG policy for learning probabilities can be expected to be ineffective as it implies too much randomness in the search when the number of measurements grows large. Therefore there is a natural question if there exist superior policies. In Section 3 we propose several such alternatives.

3. Improved policies for learning probabilities

The search for policies improving over the KG policy can be guided by the reason of its fallacy. It is a one-step-ahead decision making rule, so when taking only one measurement does not have a chance to change our decision, the KG policy will not be effective.