On the Value of Learning for Bernoulli Bandits with Unknown Parameters

Sandjai Bhulai and Ger Koole
Vrije Universiteit Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
E-mail: {sbhulai, koole}@cs.vu.nl

Published in *IEEE Transactions on Automatic Control* **45**:2135-2140, 2000

Abstract

In this paper we investigate the multi-armed bandit problem, where each arm generates an infinite sequence of Bernoulli distributed rewards. The parameters of these Bernoulli distributions are unknown and initially assumed to be Beta-distributed. Every time a bandit is selected its Beta-distribution is updated to new information in a Bayesian way. The objective is to maximize the long term discounted rewards.

We study the relationship between the necessity of acquiring additional information and the reward. This is done by considering two extreme situations which occur when a bandit has been played N times; the situation where the decision maker stops learning and the situation where the decision maker acquires full information about that bandit. We show that the difference in reward between this lower and upper bound goes to zero as N grows large.

Keywords: Bandit problem, Bayesian adaptive control, partially observed Markov decision problem.

1 Introduction

The bandit problem of Bellman is a classical problem in sequential adaptive control. The importance of this model follows from its direct applications [1, 2, 4]. Furthermore this problem is perhaps the simplest problem in the important class of Bayesian adaptive control problems [7, 8]. Even though the dynamic programming equation can be explicitly written down, it is difficult to obtain closed form solutions.

The multi-armed bandit problem of Bellman models sequential trials of alternative arms on a machine. The successive rewards for each arm forms a Bernoulli process with an unknown success probability. Since the characteristics of the processes are unknown, one learns about them when the processes are observed. This

model plays an important part in sequential clinical trials, where the arms represent alternative treatments for a disease. In this paper we model the arms as projects with unknown rewards.

There are two reasons for selecting a particular project to work on. The first reason is to obtain a high reward. The second is to acquire information which can be used to determine whether subsequent selections of the same project are profitable in future. The possible contradiction between these two reasons for choosing an arm makes the problem difficult and interesting. The amount of information one has about the projects plays an important role. It helps to answer the question whether one should select a less rewarding but more informative selection over one that is more rewarding but less informative.

A survey on bandit models can be found in Kumar [7]. Detailed expositions are given by Berry and Fristedt [2] and Gittins [4]. Gittins and Wang [5] have investigated the relationship between the importance of acquiring additional information and the amount of information which is already available. They have quantified a learning component in terms of dynamic allocation indices and have shown that this component is a decreasing function in the number of selections.

Berry and Kertz [3] have studied the worth of perfect information for multi-armed bandits. They have defined an information comparison region in order to compare the reward of the decision maker with the reward of the decision maker who has perfect information. Relations between these comparisons and the concept of regret in the minimax approach to bandit processes were established.

The methods studied in [3, 5] are cumbersome in practice, since the computational complexity is too large. In this paper we tackle this problem by adopting a direct approach. We consider two extreme situations, which occur when a bandit has been played N times; the situation where the decision maker stops learning and the situation where the decision maker acquires full information about that bandit. We express the difference in reward between this lower and upper bound in N and show that it goes to zero as N grows large.

2 Problem formulation

Suppose that there are M projects to work on. At every epoch $t = 1, 2, \ldots$ one of the projects must be selected. The work conducted on a particular project can result in either a success or a failure with a fixed unknown probability. When the conducted work is successful the project yields a reward of one unit, zero otherwise. The resulting sequence of successes and failures forms a Bernoulli process with an unknown parameter. The problem in this setting is to choose a project at each epoch such that the long term discounted rewards are maximized.

This problem can be classified as a partially observed Markov decision problem. Typically such a problem is a sequential decision process where the information concerning the parameters of interest is not available or incomplete. This feature of incompleteness cannot be ignored, e.g. by replacing uncertain parameter values with their expected values, when some actions in the system gather additional informa-

tion. Although such problems can theoretically be solved as dynamic programs, no algorithmic solution can be obtained due to an infinite state space.

The approach to solve a partially observed Markov decision problem is to transform it into an equivalent full observation problem, e.g. by techniques discussed in Rieder [10]. This method results in a Bayesian adaptive control problem, but has a state space which is very large. The state space can be reduced by choosing the Beta distribution as prior distribution for the unknown parameter. This distribution has the conjugate property, i.e., the posterior distribution belongs to the same family of distributions as the prior distribution (see DeGroot [6], Chapter 9). Since the unknown parameter lies in the interval $[0, 1]$ it is natural to choose the Beta distribution as the underlying distribution. This assumption does not pose many restrictions on our model. The Beta distribution becomes degenerate at the true value of the unknown parameter as more information is accumulated. A Beta distribution with parameters x and y will be denoted with $F_{(x,y)}$ and the corresponding probability density function with $f_{(x,y)}$. Recall that a Beta distribution $F_{(x,y)}$ with parameters $x, y \in \mathbb{N}$ has the following probability density function $f_{(x,y)}$

$$f_{(x,y)}(z) = \frac{1}{B(x,y)}\, z^{x-1}\,(1-z)^{y-1} = \frac{(x+y-1)!}{(x-1)!\,(y-1)!}\, z^{x-1}\,(1-z)^{y-1},$$

for $z \in [0, 1]$, where $B(x, y)$ is the Beta function.

The formal Markov decision problem for the multi-armed bandit can now be stated as follows. Let $\mathcal{S} = (\mathbb{N}_0 \times \mathbb{N}_0)^M$ denote the state space for the process, where state $s = (x_1, y_1, \ldots, x_M, y_M) \in \mathcal{S}$ denotes that project i is in state (x_i, y_i). The state (x, y) for a particular project represents the number of successes and the number of failures respectively which are observed for that project. Note that this state corresponds to a Beta distribution with parameters $x+1$ and $y+1$. We denote the initial state by s_1 and we assume that it is known. Let $\mathcal{A} = \{1, \ldots, M\}$ denote the action space, where action a represents selecting project a to work on. The transition probabilities are given by

$$p\big(s' \,\big|\, s, a\big) = \begin{cases} \frac{x_a+1}{x_a+y_a+2}, & \text{for } s' = s + e_{2a-1} \\ \frac{y_a+1}{x_a+y_a+2}, & \text{for } s' = s + e_{2a} \\ 0, & \text{otherwise,} \end{cases}$$

where e_i is the $2M$-dimensional unit vector with all entries zero except for the i^{th} entry, which is one. Given state s and action a the expected direct reward is determined by $F_{(x+1,y+1)}$. The expectation of such a random variable with this probability distribution yields the following expected direct reward

$$r(s, a) = \tfrac{x_a+1}{x_a+y_a+2}.$$

Then the Markov decision process is characterized by the tuple $(\mathcal{S}, \mathcal{A}, p, r)$. The set of histories at epoch t of this process is defined as the set $\mathcal{H}_t = (\mathcal{S} \times \mathcal{A})^{t-1} \times \mathcal{S}$. A

policy π is defined as the set of decision rules $(\pi_1, \pi_2, \dots)$ with $\pi_t : \mathcal{H}_t \to \mathcal{A}$. For each fixed policy π and each realization h_t of a history, the random variable A_t is given by $a_t = \pi_t(h_t)$. The random variable S_{t+1} takes values $s_{t+1} \in \mathcal{S}$ with probability $p(s_{t+1} \mid s_t, a_t)$. Let $\lambda \in (0,1)$ be the discount factor and π be a fixed policy, then the expected discounted reward criterion function $R^\pi(s_1)$ is defined by

$$R^\pi(s_1) = \mathbb{E}^\pi \sum_{t=1}^{\infty} \lambda^{t-1} r(S_t, A_t) \text{ with } S_1 = s_1.$$

Let Π denote the set of all policies. Note that $R^\pi(s_1)$ is well defined for all $\pi \in \Pi$ and $s_1 \in \mathcal{S}$, since the rewards are bounded by 1. The Markov decision problem is to find a policy π^* such that $R(s_1) = R^{\pi^*}(s_1) = \max\{R^\pi(s_1) \mid \pi \in \Pi\}$. By Theorem 6.2.10 of Puterman [9] we know that there exists an optimal deterministic stationary policy.

3 The value of information

The Markov decision model in the previous section satisfies the following dynamic programming equation

$$V(s) = \max_{i=1,\dots,M} \left\{ \tfrac{x_i+1}{x_i+y_i+2} \Big[1 + \lambda V(s + e_{2i-1}) \Big] + \tfrac{y_i+1}{x_i+y_i+2} \lambda V(s + e_{2i}) \right\},$$

where $V(s)$ denotes the optimal reward starting from state $s \in \mathcal{S}$ satisfying $s = (x_1, y_1, \dots, x_i, y_i, \dots, x_M, y_M)$. For ease of notation we denote the expression between brackets as $T_i V(s)$. Even though the dynamic programming equation can be explicitly written down, it is difficult to obtain closed form solutions or computational results because of the large state space.

The dynamic programming equation shows that the decision maker not only receives a direct reward in selecting a project, but also gains information that can lead to better decisions in future. When action a is chosen, the parameter of project a either has distribution $F_{(x_a+2,y_a+1)}$ or $F_{(x_a+1,y_a+2)}$, depending on whether a success or failure is observed. Since a random variable with such a distribution has lower variance than a random variable with probability distribution $F_{(x_a+1,y_a+1)}$ the decision maker is better informed. A formal proof of this statement is given in Lemma 3.1.

One could argue that when a particular project has been selected N times, where N can be large, enough information about the project has been obtained. Therefore basing future decisions only on this information for this project should not result in a great difference. In that case the decision maker does not need to keep record of changes in the state for this project anymore. This means that the decision maker stops learning about the unknown parameter of this particular project.

In order to compare this situation with the standard model we formulate this problem again as a Markov decision problem. Since the decision maker stops learning about a project when it has been selected N times, the corresponding state is frozen and we obtain the following *finite* state space $\underline{\mathcal{S}} = \{(x,y) \in \mathbb{N}_0 \times \mathbb{N}_0 \mid x + y \leq N\}^M$.

Take $\underline{\mathcal{A}} = \mathcal{A}$ and change the transition probabilities as follows.

$$\underline{p}\big(s' \,\big|\, s, a\big) = \begin{cases} \frac{x_a+1}{x_a+y_a+2}, & \text{for } x_a + y_a < N \text{ and } s' = s + e_{2a-1} \\ \frac{y_a+1}{x_a+y_a+2}, & \text{for } x_a + y_a < N \text{ and } s' = s + e_{2a} \\ \quad 1, & \text{for } x_a + y_a = N \text{ and } s' = s \\ \quad 0, & \text{otherwise.} \end{cases}$$

Furthermore define $\underline{r} = r$, then $(\underline{\mathcal{S}}, \underline{\mathcal{A}}, \underline{p}, \underline{r})$ defines the Markov decision process in case the decision maker stops learning about project i when the information for this project has been accumulated to N samples. The appropriate modification to the dynamic programming equation becomes

$$\underline{V}(s) = \max_{i=1,\,\ldots\,,M} \Big\{ \mathbb{1}_{\{x_i+y_i=N\}} \Big[\tfrac{x_i+1}{x_i+y_i+2} + \lambda \underline{V}(s) \Big] + \mathbb{1}_{\{x_i+y_i<N\}}\, T_i\, \underline{V}(s) \Big\}.$$

Note that in a situation where it is optimal to select an action that does not change state, that action will remain optimal. Therefore $\underline{V}(s)$ can be rewritten as follows: $\underline{V}(s) = \max \big\{ \mathbb{1}_{\{x_i+y_i=N\}} \frac{1}{1-\lambda} \frac{x_i+1}{x_i+y_i+2} + \mathbb{1}_{\{x_i+y_i<N\}}\, T_i\, \underline{V}(s) \,|\, i = 1, \ldots, M \big\}$.

Intuitively it is clear that if the decision maker does not use new information for future selections anymore, then the expected total reward will be less than $V(s)$, where this information is taken into account. The next lemma formalizes this statement.

Lemma 3.1. *$\underline{V}(s) \le V(s)$ for all $s \in \underline{\mathcal{S}}$.*

Proof. Define the operator T for functions $W : \mathcal{S} \to \mathbb{R}$ as

$$T\,W(s) = \max_{k=1,\,\ldots\,,M} \Big\{ \tfrac{x_k+1}{x_k+y_k+2} \Big[1 + \lambda W(s + e_{2k-1}) \Big] + \tfrac{y_k+1}{x_k+y_k+2}\, \lambda W(s + e_{2k}) \Big\}.$$

Define $V_0(s) = 0$ for all $s \in \mathcal{S}$ and $V_n(s) = T\,V_{n-1}(s)$. By Proposition 6.2.4 of Puterman [9] we know that the operator T is a contraction mapping on the Banach space of all bounded real valued functions on $\mathcal{S}$ endowed with the supremum norm. Therefore $V(s)$ is the unique solution to $T\,W(s) = W(s)$ and $V(s) = \lim_{n\to\infty} V_n(s)$ for arbitrary V_0.

We have to prove that a state with more information is more rewarding, therefore it suffices to prove that

$$V(s) \le \tfrac{x_i+1}{x_i+y_i+2}\, V(s + e_{2i-1}) + \tfrac{y_i+1}{x_i+y_i+2}\, V(s + e_{2i}),$$

for all $i = 1, \ldots, M$. We prove this relation by induction on n for the functions V_n. Clearly this relation holds for V_0. Fix i and suppose that the relation holds for $n \in \mathbb{N}$. Assume that the maximizing action for $V_{n+1}(s)$ is a. Since $V_{n+1}(s) = T\,V_n(s)$ we derive

$$\begin{aligned} V_{n+1}(s) &= \max_{k=1,\,\ldots\,,M} \Big\{ \tfrac{x_k+1}{x_k+y_k+2} \Big[1 + \lambda V_n(s + e_{2k-1}) \Big] + \tfrac{y_k+1}{x_k+y_k+2}\, \lambda V_n(s + e_{2k}) \Big\} \\ &= \tfrac{x_a+1}{x_a+y_a+2} \Big[1 + \lambda V_n(s + e_{2a-1}) \Big] + \tfrac{y_a+1}{x_a+y_a+2}\, \lambda V_n(s + e_{2a}). \end{aligned}$$

First suppose that $i \neq a$, then by applying the induction hypothesis we derive that the latter expression is less or equal than

$$\frac{x_a+1}{x_a+y_a+2} + \lambda \frac{x_a+1}{x_a+y_a+2}\Big[\frac{x_i+1}{x_i+y_i+2}V_n(s+e_{2a-1}+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}V_n(s+e_{2a-1}+e_{2i})\Big] + \lambda \frac{y_a+1}{x_a+y_a+2}\Big[\frac{x_i+1}{x_i+y_i+2}V_n(s+e_{2a}+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}V_n(s+e_{2a}+e_{2i})\Big].$$

By rearranging terms we find that this expression is equal to

$$\frac{x_i+1}{x_i+y_i+2}\Big[\frac{x_a+1}{x_a+y_a+2} + \lambda \frac{x_a+1}{x_a+y_a+2}V_n(s+e_{2i-1}+e_{2a-1}) + \lambda \frac{y_a+1}{x_a+y_a+2}V_n(s+e_{2i-1}+e_{2a})\Big] + \frac{y_i+1}{x_i+y_i+2}\Big[\frac{x_a+1}{x_a+y_a+2} + \lambda \frac{x_a+1}{x_a+y_a+2}V_n(s+e_{2i}+e_{2a-1}) + \lambda \frac{y_a+1}{x_a+y_a+2}V_n(s+e_{2i}+e_{2a})\Big].$$

By definition of T_a, the latter expression is equal to

$$\frac{x_i+1}{x_i+y_i+2}\,T_aV_n(s+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}\,T_aV_n(s+e_{2i})$$

$$\leq \max_{k=1,\dots,M}\Big\{\frac{x_i+1}{x_i+y_i+2}\,T_kV_n(s+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}\,T_kV_n(s+e_{2i})\Big\}.$$

The argument for $i = a$ is completely analogous. Hence we have

$$\begin{aligned} V_{n+1}(s) &= \max_{k=1,\dots,M}\Big\{\frac{x_k+1}{x_k+y_k+2}\Big[1+\lambda V_n(s+e_{2k-1})\Big] + \frac{y_k+1}{x_k+y_k+2}\,\lambda V_n(s+e_{2k})\Big\} \\ &\leq \max_{k=1,\dots,M}\Big\{\frac{x_i+1}{x_i+y_i+2}\,T_kV_n(s+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}\,T_kV_n(s+e_{2i})\Big\} \\ &\leq \frac{x_i+1}{x_i+y_i+2}\max_{k=1,\dots,M}\big\{T_kV_n(s+e_{2i-1})\big\} + \frac{y_i+1}{x_i+y_i+2}\max_{k=1,\dots,M}\big\{T_kV_n(s+e_{2i})\big\} \\ &= \frac{x_i+1}{x_i+y_i+2}V_{n+1}(s+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2}\,V_{n+1}(s+e_{2i}). \end{aligned}$$

The proof is concluded by taking the limit in n. □

At this point we know that $\underline{R}(s) \leq R(s)$ and our main interest is to compare $\underline{R}(s)$ with $R(s)$, i.e., to give an upper bound to $R(s) - \underline{R}(s)$. However, it is not straightforward to carry out this computation, therefore we define another Markov decision process which has a higher reward than $R(s)$ and which will facilitate the comparison.

Suppose that when a particular project has been selected N times, the decision maker obtains full information about the unknown parameter of that project. Then the decision maker does not need to learn anything about the unknown parameter and can base his future decisions on the realization of the unknown parameter.

The state space of the process is equal to $\underline{\mathcal{S}}$ when none of the projects have been selected N times. After this moment the state space changes to a single value representing the realization of the unknown parameter. This results in a complicated state space. The state space can be represented by $\overline{\mathcal{S}} = \big\{\overline{s} \mid \overline{s} = (s,z) \in \underline{\mathcal{S}} \times [0,1]^M\big\}$. In this case the state $\overline{s}$ consists of s with augmented extra information z. The value

of z_i represents the realization of the unknown parameter of project i. Although this information can only be used after a project has been selected N times, we define the state space slightly bigger for ease of notation.

Let $\overline{V}(\overline{s})$ denote the optimal expected reward starting in state $\overline{s} \in \overline{\mathcal{S}}$. Note that the set of states which V and $\overline{V}$ share is given by $\underline{\mathcal{S}}$. For states $s \in \underline{\mathcal{S}}$ the extra information z is not available and can be disregarded. Therefore we will write $\overline{V}(s)$ instead of $\overline{V}\big((s,z)\big)$ for all $s \in \underline{\mathcal{S}}$ and $z \in [0,1]$. Intuitively it is clear that if the decision maker has full information, then the expected total reward will be greater than $V(s)$ for all $s \in \underline{\mathcal{S}}$. The following lemma justifies this intuition.

Lemma 3.2. *$V(s) \leq \overline{V}(s)$ for all $s \in \underline{\mathcal{S}}$.*

Proof. Let $s \in \underline{\mathcal{S}}$ and fix $i \in \mathcal{A}$. Define $\overline{V}_0^i(s) = V(s)$ and $\overline{V}_n^i(s)$ inductively by

$$\overline{V}_n^i(s) = \tfrac{x_i+1}{x_i+y_i+2}\,\overline{V}_{n-1}^i(s+e_{2i-1}) + \tfrac{y_i+1}{x_i+y_i+2}\,\overline{V}_{n-1}^i(s+e_{2i}).$$

Note that $\overline{V}_n^i(s)$ represents the situation where the decision maker has already acquired more information about project i. The decision maker looks n steps ahead and consequently knows more about project i. In Lemma 3.1 it was proven that $V(s) \leq \frac{x_i+1}{x_i+y_i+2} V(s+e_{2i-1}) + \frac{y_i+1}{x_i+y_i+2} V(s+e_{2i})$, for all $i \in \mathcal{A}$. By repeating this argument we derive $0 \leq \overline{V}_0^i(s) \leq \overline{V}_1^i(s) \leq \cdots$. Now define $\overline{V}_n(s)$ as follows

$$\overline{V}_n(s) = \max_{i=1,\ldots,M} \Big\{ \mathbb{1}_{\{x_i+y_i=N\}} \overline{V}_n^i(s) + \mathbb{1}_{\{x_i+y_i<N\}}\, T_i\, V(s) \Big\}.$$

Then it follows that $V(s) \leq \overline{V}_n(s)$ for all $n \in \mathbb{N}$. Note that $\overline{V}(s) = \lim_{n\to\infty} \overline{V}_n(s)$; the case where the decision maker has full information. By the Monotone Convergence Theorem it finally follows that $V(s) \leq \overline{V}(s)$. □

Comparing $\overline{R}(s)$ with $\underline{R}(s)$ for a given $s \in \underline{\mathcal{S}}$ is still not easy, since both processes differ in the amount of information when a project has already been selected N times. In case of $\overline{R}(s)$ the decision maker knows $\overline{\mathcal{S}}$ which has extra information z. However in case of $\underline{R}(s)$ the decision maker has less information, because he only knows $\underline{\mathcal{S}}$. Therefore for a given a policy $\overline{\pi} \in \overline{\Pi}$, with decision rules based on z, one cannot compare $\overline{R}^{\overline{\pi}}(s) - \underline{R}^{\overline{\pi}}(s)$, since the latter term is not well defined.

The value of the realization z is determined by the probability distribution of the unknown parameter. If we adjust this probability distribution, then we will be able to carry out the comparison. Let $\overline{\overline{F}}_{(x+1,y+1)}$ denote the probability distribution with positive probability mass concentrated on only two points as follows. Let $\beta = \frac{x+1}{x+y+2}$ and choose $0 < \delta < \frac{1}{N+2}$, then $\beta + \delta < 1$. Define the probability mass function $\overline{\overline{f}}$ by

$$\overline{\overline{f}}_{(x+1,y+1)}(z) = \begin{cases} \int_0^{\beta+\delta} f_{(x+1,y+1)}(u)\,\mathrm{d}u & \text{for } z = \beta+\delta \\ \int_{\beta+\delta}^1 f_{(x+1,y+1)}(u)\,\mathrm{d}u & \text{for } z = 1. \end{cases}$$

The Markov decision process $(\overline{\overline{\mathcal{S}}}, \overline{\overline{\mathcal{A}}}, \overline{\overline{p}}, \overline{\overline{r}})$ can now be defined as follows. Let $\overline{\overline{\mathcal{S}}} = \overline{\mathcal{S}}$. Define the action space $\overline{\overline{\mathcal{A}}} = \mathcal{A}$ and the transition probabilities by

$$\overline{\overline{p}}\big(\overline{\overline{s}}' \,\big|\, \overline{\overline{s}}, a\big) = \begin{cases} \frac{x_a+1}{x_a+y_a+2}, & \text{for } x_a+y_a < N-1 \text{ and } s' = s+e_{2a-1} \\ \frac{y_a+1}{x_a+y_a+2}, & \text{for } x_a+y_a < N-1 \text{ and } s' = s+e_{2a} \\ \frac{x_a+1}{x_a+y_a+2}\,\overline{\overline{f}}_{(x_a+2,y_a+1)}(u) & \text{for } x_a+y_a = N-1, s' = s+e_{2a-1} \text{ and } z'_a = u \\ \frac{y_a+1}{x_a+y_a+2}\,\overline{\overline{f}}_{(x_a+1,y_a+2)}(u) & \text{for } x_a+y_a = N-1, s' = s+e_{2a} \text{ and } z'_a = u \\ 1, & \text{for } x_a+y_a = N \text{ and } \overline{s}' = \overline{s} \\ 0, & \text{otherwise,} \end{cases}$$

where $u \in [0,1]$. Define the direct reward by

$$\overline{\overline{r}}\big((s,z),a\big) = \begin{cases} \frac{x_a+1}{x_a+y_a+2}, & \text{for } x_a+y_a < N \\ z_a, & \text{for } x_a+y_a = N. \end{cases}$$

Note that the difference between $\overline{V}$ and $\overline{\overline{V}}$ is reflected in the transition probabilities. In the former case one had to deal with a continuous probability distribution. In the latter case the probability distribution is discrete and concentrated on two special points only. If Z_1 and Z_2 are two random variables with probability distribution $F_{(x,y)}$ and $\overline{\overline{F}}_{(x,y)}$ respectively, then Z_2 is stochastically larger than Z_1; i.e., $\mathbb{P}(Z_1 > z) \leq \mathbb{P}(Z_2 > z)$ for all $z \in [0,1]$. From Proposition 8.1.2, Ross [11] we know that in this case $\mathbb{E}[h(Z_1)] \leq \mathbb{E}[h(Z_2)]$ for all increasing functions h. Therefore we have the following corollary.

Corollary 3.3. *$0 \leq \underline{R}(s) \leq R(s) \leq \overline{R}(s) \leq \overline{\overline{R}}(s)$ for all $s \in \underline{\mathcal{S}}$.*

The process with reward $\overline{\overline{R}}(s)$ is constructed in such a way, that the information structure when a project has been selected N times is nearly the same as in $\underline{R}(s)$. The decision maker either observes $\beta + \delta$ or 1 as the realization for the unknown parameter. In the first case the decision maker has the same information as in $\underline{R}(s)$, namely the expectation. In the second case we know that since 1 is the highest possible reward, the decision maker is going to select that project continuously in future. This fact enables us to prove the main theorem.

Theorem 3.4. *$0 \leq R(s) - \underline{R}(s) \leq \max_{i \in \mathcal{A}} \frac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\Big[\delta + \frac{l(s)}{\delta^2\,(N+3)}\Big]$ for all $s \in \underline{\mathcal{S}}$ and $\delta < \frac{1}{N+2}$ where $l(s) = \sum_{i=1}^{M} \mathbb{1}_{\{x_i+y_i<N\}}$.*

Proof. Because of Lemma 3.1 the difference $R(s) - \underline{R}(s)$ is non-negative. Also by Corollary 3.3 we know that $R(s) - \underline{R}(s) \leq \overline{\overline{R}}(s) - \underline{R}(s)$. Therefore it suffices to prove the bound for the latter term. We adopt the same approach as in the proof of Lemma 3.1. Define the operator $\underline{T}$ for functions $W : \underline{\mathcal{S}} \to \mathbb{R}$ as

$$\underline{T}\,W(s) = \max_{i=1,\ldots,M} \Big\{ \mathbb{1}_{\{x_i+y_i=N\}} \Big[\tfrac{1}{1-\lambda}\, \tfrac{x_i+1}{x_i+y_i+2} \Big] + \mathbb{1}_{\{x_i+y_i<N\}}\, T_i\, W(s) \Big\}.$$

Define $\underline{V}_0(s) = 0$ for all $s \in \underline{\mathcal{S}}$ and $\underline{V}_n(s) = \underline{T}\,\underline{V}_{n-1}(s)$. By Proposition 6.2.4 of Puterman [9] we know that the operator $\underline{T}$ is a contraction mapping on the Banach space of all bounded real valued functions on $\underline{\mathcal{S}}$ endowed with the supremum norm. Therefore $\underline{V}(s)$ is the unique solution to $\underline{T}\,W(s) = W(s)$ and $\underline{V}(s) = \lim_{n\to\infty} \underline{V}_n(s)$ for arbitrary $\underline{V}_0$. Similarly, the operator $\overline{\overline{T}}$ is defined for functions $W : \overline{\overline{\mathcal{S}}} \to \mathbb{R}$ as

$$\begin{aligned}
\overline{\overline{T}}\,W(s) = \max_{i=1,\dots,M} \Bigg\{ & \mathbb{1}_{\{x_i+y_i=N\}} \tfrac{1}{1-\lambda}\Big(\tfrac{x_i+1}{x_i+y_i+2} + \delta\Big) + \\
& \mathbb{1}_{\{x_i+y_i=N-1\}} \Bigg[\tfrac{x_i+1}{x_i+y_i+2} + \lambda \tfrac{x_i+1}{x_i+y_i+2}\Big(q_{(x_i+1,y_i)} W(s+e_{2i-1}) + \big(1-q_{(x_i+1,y_i)}\big)\tfrac{1}{1-\lambda}\Big) + \\
& \lambda \tfrac{y_i+1}{x_i+y_i+2}\Big(q_{(x_i,y_i+1)} W(s+e_{2i}) + \big(1-q_{(x_i,y_i+1)}\big)\tfrac{1}{1-\lambda}\Big)\Bigg] + \mathbb{1}_{\{x_i+y_i<N-1\}}\, T_i\, W(s) \Bigg\},
\end{aligned}$$

where $q_{(x_i,y_i)} = \int_0^{\beta+\delta} f_{(x_i+1,y_i+1)}(z)\,\mathrm{d}z$ with $\beta = \frac{x_i+1}{x_i+y_i+2}$. Note that this equation represents the situation where the decision maker receives the realization of the unknown parameter (under the modified probability distribution) after selecting a project $N-1$ times. When it is optimal to select this project again after this moment, then it will be optimal to select it continuously thereafter, since the state does not change.

Now we prove the statement by induction. Let $s \in \underline{\mathcal{S}}$, then clearly the statement holds for $\overline{\overline{V}}_0(s) - \underline{V}_0(s) = 0$. Now suppose that the statement holds for $n \in \mathbb{N}$. Assume w.l.o.g. that the first m projects have reached level N, the second m' projects level $N-1$ and the remaining $M-m-m'$ projects have not reached level $N-1$ yet for an arbitrary fixed $m \in \{0, \dots, M\}$ and $m' \in \{0, \dots, M-m\}$. Now assume that it is optimal to choose one of the first m projects. Then for $i = 1, \dots, m$ the difference is less than

$$\Big[\tfrac{1}{1-\lambda}\Big(\tfrac{x_i+1}{x_i+y_i+2}+\delta\Big)\Big] - \Big[\tfrac{1}{1-\lambda}\,\tfrac{x_i+1}{x_i+y_i+2}\Big] = \tfrac{\delta\,\lambda^{N-(x_i+y_i)}}{1-\lambda} \le \max_{k\in\mathcal{A}} \tfrac{\lambda^{N-(x_k+y_k)}}{1-\lambda}\Big[\delta + \tfrac{l(s)}{\delta^2\,(N+3)}\Big].$$

Next consider the second m' projects. First note that for a random variable U with a Beta distribution with parameters (x_i+1, y_i+1) the following holds.

$$\begin{aligned}
1 - q_{(x_i,y_i)} &= \mathbb{P}\big(U > \beta+\delta\big) \le \mathbb{P}\big(\{U < \beta-\delta\} \cup \{U > \beta+\delta\}\big) \\
&= \mathbb{P}\big(\{U-\beta < -\delta\} \cup \{U-\beta > \delta\}\big) = \mathbb{P}\big(|U-\beta| > \delta\big) \\
&\le \tfrac{\mathbb{V}\mathrm{ar}\,U}{\delta^2} \le \tfrac{1}{\delta^2\,(x_i+y_i+3)}.
\end{aligned}$$

The last inequality follows by Chebyshev's Inequality. Now for $j = m+1, \dots, m+m'$ the difference is given by

$$\begin{aligned}
& \Big[\tfrac{x_j+1}{x_j+y_j+2} + \lambda\,\tfrac{x_j+1}{x_j+y_j+2}\Big(q_{(x_j+1,y_j)}\,\overline{\overline{V}}_n(s+e_{2j-1}) + \big(1-q_{(x_j+1,y_j)}\big)\tfrac{1}{1-\lambda}\Big) + \\
& \lambda\,\tfrac{y_j+1}{x_j+y_j+2}\Big(q_{(x_j,y_j+1)}\,\overline{\overline{V}}_n(s+e_{2j}) + \big(1-q_{(x_j,y_j+1)}\big)\tfrac{1}{1-\lambda}\Big)\Big] - T_j\,\underline{V}_n(s)
\end{aligned}$$

$$\leq \lambda \tfrac{x_j+1}{x_j+y_j+2}\Big(q_{(x_j+1,y_j)}\overline{\overline{V}}_n(s+e_{2j-1}) + \big(1-q_{(x_j+1,y_j)}\big)\tfrac{1}{1-\lambda} - \underline{V}_n(s+e_{2j-1})\Big) +$$
$$\lambda \tfrac{y_j+1}{x_j+y_j+2}\Big(q_{(x_j,y_j+1)}\overline{\overline{V}}_n(s+e_{2j}) + \big(1-q_{(x_j,y_j+1)}\big)\tfrac{1}{1-\lambda} - \underline{V}_n(s+e_{2j})\Big)$$

$$\leq \lambda \tfrac{x_j+1}{x_j+y_j+2}\Big(\Big[\overline{\overline{V}}_n(s+e_{2j-1}) - \underline{V}_n(s+e_{2j-1})\Big] + \tfrac{1}{\delta^2\,(N+3)}\,\tfrac{1}{1-\lambda}\Big) +$$
$$\lambda \tfrac{y_j+1}{x_j+y_j+2}\Big(\Big[\overline{\overline{V}}_n(s+e_{2j}) - \underline{V}_n(s+e_{2j})\Big] + \tfrac{1}{\delta^2\,(N+3)}\,\tfrac{1}{1-\lambda}\Big).$$

Note that $l(s+e_{2j-1}) = l(s+e_{2j}) = l(s)-1$. By applying the induction hypothesis we derive that the last expression is less than

$$\tfrac{\lambda}{1-\lambda}\,\Big(\max\Big\{\lambda^{N-(x_i+y_i)}; i\in\mathcal{A}\setminus\{j\},\, \lambda^{N-(x_j+y_j+1)}\Big\}\Big[\delta + \tfrac{l(s)-1}{\delta^2\,(N+3)}\Big] + \tfrac{1}{\delta^2\,(N+3)}\Big)$$
$$\leq \max_{i\in\mathcal{A}}\,\tfrac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\,\Big[\delta + \tfrac{l(s)}{\delta^2\,(N+3)}\Big].$$

Finally consider the last $M-m-m'$ projects. Note that for $k = m+m', \ldots, M$ we have $l(s) = l(s+e_{2k-1}) = l(s+e_{2k})$. Therefore the expression $T_k\left[\overline{\overline{V}}_n(s) - \underline{V}_n(s)\right]$ is given by

$$\tfrac{x_k+1}{x_k+y_k+2}\,\lambda\Big(\overline{\overline{V}}_n(s+e_{2k-1}) - \underline{V}_n(s+e_{2k-1})\Big) + \tfrac{y_k+1}{x_k+y_k+2}\,\lambda\Big(\overline{\overline{V}}_n(s+e_{2k}) - \underline{V}_n(s+e_{2k})\Big)$$
$$\leq \tfrac{\lambda}{1-\lambda}\,\max\Big\{\lambda^{N-(x_i+y_i)}; i\in\mathcal{A}\setminus\{j\},\, \lambda^{N-(x_j+y_j+1)}\Big\}\Big[\delta + \tfrac{1}{\delta^2\,(N+3)}\Big]$$
$$\leq \max_{i\in\mathcal{A}}\,\tfrac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\,\Big[\delta + \tfrac{1}{\delta^2\,(N+3)}\Big].$$

Now it follows that $\overline{\overline{V}}_{n+1}(s) - \underline{V}_{n+1}(s)$ satisfies the statement of the theorem. The proof is concluded by taking the limit in n. □

The bounds in the previous theorem still contain $\delta > 0$. Since δ was arbitrarily chosen, we can minimize the bound for fixed N with respect to δ. This will result in a bound independent of δ and the result is stated in the following theorem.

Theorem 3.5. $0 \leq R(s) - \underline{R}(s) \leq \max_{i\in\mathcal{A}}\,\tfrac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\,\tfrac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}}\,\sqrt[3]{l(s)}$ *for all* $s\in\underline{\mathcal{S}}$.

Proof. Let N be fixed and define $g(\delta) = \max_{i\in\mathcal{A}}\,\tfrac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\,\Big[\delta + \tfrac{l(s)}{\delta^2\,(N+3)}\Big]$. Then

$$\hat{\delta} = \sqrt[3]{\tfrac{2\,l(s)}{N+3}} \quad \text{solves} \quad \frac{\mathrm{d}\,g(\delta)}{\mathrm{d}\,\delta} = \max_{i\in\mathcal{A}}\,\tfrac{\lambda^{N-(x_i+y_i)}}{1-\lambda}\,\Big[1 - \tfrac{2\,l(s)}{\delta^3\,(N+3)}\Big] = 0.$$

Hence $\hat{\delta}$ minimizes g. Although Theorem 3.4 is formulated for $\delta < \frac{1}{N+2}$ the theorem holds for general $\delta > 0$. When $\overline{\overline{f}}$ is defined to be degenerate at 1 when $\beta+\delta \geq 1$, one can easily check that Theorem 3.4 still holds. Now the theorem follows by substituting $\hat{\delta}$ in g, since $0 \leq R(s) - \underline{R}(s) \leq g(\hat{\delta})$. □

Observe that the bound in Theorem 3.5 has the property that the difference goes to zero as N grows large. However, this is not due to discounting, since

$$R(s) - \underline{R}(s) \leq \max_{i \in \mathcal{A}} \frac{\lambda^{N-(x_i+y_i)}}{1-\lambda} \frac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}} \sqrt[3]{l(s)} \leq \frac{1}{1-\lambda} \frac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}} \sqrt[3]{M}.$$

The latter bound, which is less tight, also goes to zero as N grows large even without the discount factor. Moreover, observe that this bound also holds for any state. One can even show that the bound for the reward at any time t is $\left[(\sqrt[3]{2} + \sqrt[3]{\frac{1}{4}})\sqrt[3]{M}\right] / \sqrt[3]{N+3}$.

4 Numerical results

In this section we illustrate Theorem 3.5 derived in the previous section by showing that in practice the state space can indeed be chosen finite in order to be close to the optimal solution.

Suppose that in the initial state $s_1 = (0, \ldots, 0)$ the decision maker wants to obtain a solution which differs less than $\varepsilon = 10^{-3}$ from the optimal solution. We call such a solution a ε-optimal solution. Note that the initial state s_1 represents the situation where the decision maker does not have any information about the unknown parameters of the projects.

By using the bounds derived in Theorem 3.5 one can determine the value of N for which the decision maker can stop learning about the unknown parameter of a particular project. However since the total reward will increase for λ close to one, the value of N will grow large. Therefore it is better to look at the relative difference $\frac{R(s)-\underline{R}(s)}{R(s)}$. This leads to the following table when M and λ are varied.

λ	$M = 2$	$M = 3$
0.7	21	21
0.8	32	33
0.9	66	66

In practice it suffices to take smaller values of N. Figure 1 depicts two situations with two and three projects respectively. The three lines reflect the cases with $\lambda = 0.90$, $\lambda = 0.80$ and $\lambda = 0.70$ from top to down respectively for $\underline{V}(s_1)$. The dashed lines represent the bounds on the error on the total reward obtained by using $\underline{V}(s_1)$ instead of $V(s_1)$; thus the dashed lines represent the upperbound $V(s_1) + \left[\overline{\overline{V}}(s_1) - \underline{V}(s)\right]$.

One can see that the total reward converges very fast already for small values of N. It turns out that for $\lambda = 0.90$ one can take $N = 28$ instead of $N = 108$ in order to derive a ϵ-optimal total reward. The following table summarizes the values of N which can be taken instead of the larger value which one derives from the theorem.

λ	$M = 2$	$M = 3$
0.7	9	10
0.8	13	14
0.9	28	33

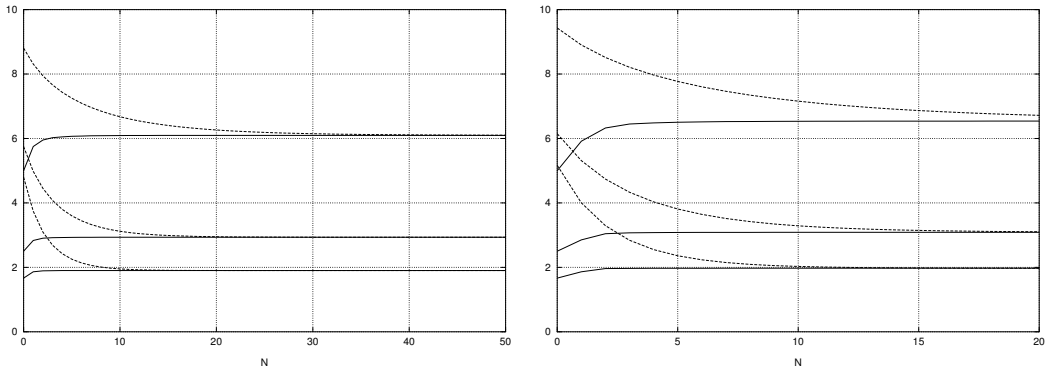

Figure 1: Comparison of bounds

These values of N are small enough to make the problem computationally tractable and to derive a ε-optimal solution.

References

[1] R. Bellman. "A problem in the sequential design of experiments". *Sankhya*, 16:221–229, 1956.

[2] D.A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

[3] D.A. Berry and R.P. Kertz. "Worth of perfect information in Bernoulli bandits". *Advances in Applied Probability*, 23:1–23, 1991.

[4] J.C. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 1989.

[5] J. Gittins and Y. Wang. "The learning component of dynamic allocation indices". *The Annals of Statistics*, 20:1625–1636, 1992.

[6] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[7] P.R. Kumar. "A survey of some results in stochastic adaptive control". *SIAM Journal of Control and Optimization*, 23:329–380, 1985.

[8] P.R. Kumar and T.I. Seidman. "On the optimal solution of the one-armed bandit adaptive control problem". *IEEE Transactions on Automatic Control*, 26:1176–1184, 1981.

[9] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[10] U. Rieder. "Bayesian dynamic programming". *Advances in Applied Probability*, 7:330–348, 1975.

[11] S.M. Ross. *Stochastic Processes*. John Wiley & Sons, 1983.