
A SAMPLING-BASED GITTINS INDEX APPROXIMATION

A PREPRINT

Stef Baas

Stochastic Operations Research Group, University of Twente, Enschede, The Netherlands
s.p.r.baas@utwente.nl

Richard J. Boucherie

Stochastic Operations Research Group, University of Twente, Enschede, The Netherlands

Aleida Braaksma

Stochastic Operations Research Group, University of Twente, Enschede, The Netherlands

July 24, 2023

ABSTRACT

A sampling-based method is introduced to approximate the Gittins index for a general family of alternative bandit processes. The approximation consists of a truncation of the optimization horizon and support for the immediate rewards, an optimal stopping value approximation, and a stochastic approximation procedure. Finite-time error bounds are given for the three approximations, leading to a procedure to construct a confidence interval for the Gittins index using a finite number of Monte Carlo samples, as well as an epsilon-optimal policy for the Bayesian multi-armed bandit. Proofs are given for almost sure convergence and convergence in distribution for the sampling based Gittins index approximation. In a numerical study, the approximation quality of the proposed method is verified for the Bernoulli bandit and Gaussian bandit with known variance, and the method is shown to significantly outperform Thompson sampling and the Bayesian Upper Confidence Bound algorithms for a novel random effects multi-armed bandit.

Keywords Stochastic Approximation · Multi-Armed Bandits · Optimal Stopping · Bayesian Computation · Markov Decision Processes

1 Introduction

The family of alternative bandit processes (FABP) is a well established problem in the field of Operations Research (Glazebrook [1983]). In short, an FABP is a problem in which the decision maker sequentially chooses one out of a finite collection of independent Markov reward processes (sometimes called arms) to evolve, and the goal is to optimize the total discounted reward. As initially proven in Gittins [1979], an index value can be determined based on the state of each bandit process, and the optimal policy is to choose the arm with the highest index value at each decision epoch. This index value, introduced as the dynamic allocation index, is now referred to as the Gittins index. As the Gittins index can be calculated separately based on the state of each bandit process, there is a large gain in computational efficiency for finding the optimal policy when compared to methods taking the states of all bandit processes into account (Puterman [1990]). Next to the family of alternative bandit processes, the Gittins index was also found to be the optimal solution to a number of other problems in operations research such as optimal scheduling and search problems (Aalto et al. [2011], Boodaghians et al. [2023]).

When, e.g., the total average reward is considered, or the Markov reward processes are no longer independent, the bandit problem becomes a restless bandit problem, where each arm evolves after an arm is chosen, instead of just that specific arm. Optimality of the Gittins index policy is no longer a guarantee in this case. When the condition of indexability is met, however, the policy choosing the largest Whittle index can provide a well-behaving heuristic for the problem under

consideration (Weber and Weiss [1990], Glazebrook and Minty [2009], ?). The Whittle index has a definition similar to the Gittins index, and hence computation methods for the Gittins index work for the Whittle index as well.

Families of alternative bandit processes arise in various applications, such as job scheduling in queues, mining (or search) operations, advertisement, and Bayesian multi-armed bandit problems (Mahajan and Teneketzis [2008], Gittins et al. [2011]). The current paper will focus mainly on multi-armed bandit problems, with an application to clinical trials. Due to an increase in computational power and an increased focus on patient-centric medicine, research on treatment allocation in clinical trials using the Gittins index has gained popularity in the last decade (Villar et al. [2015a,b], Robertson et al. [2023]). In the multi-armed bandit problem, the decision maker is tasked with the choice to sample rewards from one out of a finite collection of unknown distributions at each decision epoch. In the Bayesian setting, each distribution is endowed with a prior. The resulting sequences of posterior distributions and posterior mean outcomes at each decision epoch then result in the separate Markov reward processes of an FABP. The Bayesian multi-armed bandit problem, popularized in Robbins [1952], was introduced in Thompson [1933] along with the approximate solution method now referred to as Thompson sampling, which was hence the first approximate solution method for the Bayesian multi-armed bandit problem (Slivkins [2019]).

This paper focuses on maximizing the Bayesian total expected discounted reward. When the regret under a multi-armed bandit problem is analyzed in the frequentist framework, many index-based approximate solution methods exist (Bubeck and Cesa-Bianchi [2012], Kuleshov and Precup [2014], Lattimore and Szepesvári [2020]), either based on frequentist or Bayesian approaches. In Lai et al. [1985], frequentist asymptotic lower bounds were found for the number of suboptimal choices made under a broad class of approximate solution methods. In the case of bounded rewards, the asymptotic regret upper bound equals this lower bound for e.g., approximate solution methods based on an upper confidence bound (UCB) for the expected reward (Lai [1987], Auer et al. [2002]), proving that these methods are asymptotically optimal in the frequentist framework. Many Bayesian approximate solution methods are asymptotically optimal and were shown to have excellent empirical frequentist performance (Kaufmann et al. [2012a], Kaufmann [2018], Lattimore and Szepesvári [2020]). In Lattimore [2016] it was shown that, in the frequentist framework, the Gittins index strategy with improper uniform prior for the mean yields an asymptotically optimal approximate solution method for the Gaussian model, outperforming other optimal Bayesian methods such as the Thompson sampling and Bayesian Upper Confidence Bound (Bayes-UCB).

There is a large amount of literature on calculating the Gittins index, starting with the Calibration method introduced in Gittins [1979]. A survey covering most literature up to 2014 on offline and online algorithms to calculate the Gittins index for general countable state FABPs is provided in Chakravorty and Mahajan [2014]. In Yao [2006], an approximation method for the Gittins index is derived in the Gaussian bandit with unknown mean and known variance. A numerical approximation for this bandit based on quadratic splines is described in Lattimore [2016]. In Edwards [2019] the lack of general open source code for calculating the Gittins index is acknowledged, and methods and accompanying code, based on the Calibration method in Gittins [1979], are given to calculate the Gittins index for the Bernoulli and Gaussian bandit with known variance.

Calculation methods found in literature only work for countable state FABPs or are tailored to the Gaussian bandit with known variance. The calculation methods often revolve around the Bellman equation where it is assumed that the transition probabilities are known in closed form. First, when dealing with experimental data, the models can be much more complex and the assumption of known transition probabilities might not hold. For instance, the main types of outcome data encountered in clinical trials are categorical, continuous or event-times. Clearly only the first of these three is covered by considering a finite or countable state space, as continuous outcome data are often modeled using parameters on an uncountable parameter space. Second, in, e.g., latent variable modeling, the posterior distribution of the model parameters is often not known in closed form. Markov chain Monte Carlo approaches (Gilks et al. [1995]) provide a means to do (approximate) posterior inference in this case. Hence, the transition structure of the FABP is unknown, however one can still sample (approximately) from the Markov reward process. Another situation where the posterior is not available in closed form is when the prior being used in the Bayesian analysis is nonconjugate with respect to the likelihood. For instance, when a Bayesian experiment is performed for Gaussian data with possibly conflicting prior information, a nonconjugate Student's t -distribution can be assumed as a prior distribution for the mean (Neuenschwander et al. [2020]). No method exists to accurately approximate the Gittins index in these settings.

To address the open problems listed above, a sampling-based Gittins index approximation (SBGIA) is introduced in this paper. The SBGIA can be calculated for any type of state space, also when there is access only to a simulator for future rewards. We first approximate the Gittins index using a truncation of the optimization horizon and support for the immediate rewards of the Markov chains. Second, we use a stochastic approximation procedure (Robbins and Monro [1951]) based on the optimal stopping value approximation introduced in Chen and Goldberg [2018] to find the root (i.e., the fair charge) of the prevailing charge formulation of the Gittins index, see, e.g., Weber [1992, Equation 2]. As the SBGIA

is sampling-based, samples from the (approximate) posterior reward distribution can be used to make decisions in the proposed algorithm.

The paper is organized as follows. Section 2 introduces the family of alternative bandit processes. Section 3 extends optimal stopping value approximation results from Chen and Goldberg [2018] to reward processes that are not restricted to be non-negative. In Section 4.1, based on convergence results provided in Chen and Goldberg [2018], we obtain finite-time convergence results for the SBGIA. In Section 4.2 we prove asymptotic convergence results for the stochastic approximation iterates. In Section 5 we show the performance of the SBGIA in several numerical simulation studies. Appendix A states the longer proofs of the theorems in the paper, and Appendix B summarises the notation used in the paper.

2 Family of alternative bandit processes

We consider A independent Markov chains $(S_t^a)_t$ for $a \in [A] = \{1, \dots, A\}$, referred to as *arms*, each on a (shared) Borel space $(\mathcal{S}, \mathcal{G})$ with underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The (shared) transition kernel is denoted \mathbb{P}_s , where s denotes the initial state of the Markov chain, and expectations w.r.t. \mathbb{P}_s are denoted \mathbb{E}_s . Let R denote a $\mathcal{G}/\mathcal{B}(\mathbb{R})$ -measurable reward function. See Lattimore and Szepesvári [2020, Chapter 35] for details on this setting.

Let \mathcal{P} be the set of policies, i.e., the mappings from the set of histories

$$\mathcal{H} = \{(s_u^1, \dots, s_u^A, a_u)_{u=0}^{t-1} \times (s_t^1, \dots, s_t^A) : t \in \mathbb{N}_0, a_u \in [A], s_u^a \in \mathcal{S} \quad \forall a \in [A], u \in \mathbb{N}_0\} \quad (1)$$

to the unit- A -simplex Δ^A of probability vectors over $[A]$. A fixed policy $\pi \in \mathcal{P}$ induces a Markov chain $\left(\left((S_u^a)_{u=1}^{N_{a,t}^\pi} \right)_{a \in [A]} \right)_t$, where $N_{a,t}^\pi$ is the number of times arm a is chosen by the policy π up to and including time t . We denote the probability measure and expectation under this fixed policy by \mathbb{P}_π and \mathbb{E}_π . The objective is to find the optimal policy maximizing the total sum of discounted rewards for a discount factor $\gamma \in (0, 1)$:

$$\pi^* = \arg \max_{\pi \in \mathcal{P}} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_{N_{A_t, t}^a}^a) \right]. \quad (2)$$

We furthermore assume that the reward function for arm a , with initial state s^a , is discounted absolutely convergent in expectation under discount factor γ

Assumption 1.

$$C(s^a) = \mathbb{E}_{s^a} \left[\sum_{t=0}^{\infty} \gamma^t |R(S_t^a)| \right] < \infty \quad \forall s^a \in \mathcal{S}.$$

Under Assumption 1, the $\arg \max$ in (2) is attained, and the optimal policy π^* for the Markov decision process above is the policy choosing the arm with the highest Gittins index, see, e.g., Lattimore and Szepesvári [2020, Theorem 35.9]. To specify this, let s_h^a be the current state for arm a in history $h \in \mathcal{H}$, then

$$\pi^*(h) \in \arg \max_{a \in [A]} \nu(s_h^a), \quad \text{with} \quad \nu(s^a) = \sup_{\tau \in \mathcal{T}^a} \frac{\mathbb{E}_{s^a} \left[\sum_{t=0}^{\tau-1} \gamma^t R(S_t^a) \right]}{\mathbb{E}_{s^a} \left[\sum_{t=0}^{\tau-1} \gamma^t \right]} \quad \forall s^a \in \mathcal{S}, \quad (3)$$

where \mathcal{T}^a is the set of stopping times in \mathbb{N} w.r.t. the filtration $(\mathcal{F}_t^a)_t$ generated by the process $(S_t^a)_t$ starting from state s^a . The *Gittins index* $\nu(s_h^a)$ hence only depends on the current state s_h^a for any arm a . When ties occur in the above expression, i.e., the $\arg \max$ returns more than one value, the policy uniformly chooses an arm from the different choices of $\arg \max$. Hence, under this choice, the Gittins index policy is a randomized Markov policy.

The Gittins index can also be written as (Weber [1992, Equation 2])

$$\nu(s^a) = \sup \left\{ \nu : \sup_{\tau \in \mathcal{T}^a} \mathbb{E}_{s^a} \left[\sum_{t=0}^{\tau-1} \gamma^t (R(S_t^a) - \nu) \right] \geq 0 \right\} \quad \forall s^a \in \mathcal{S}. \quad (4)$$

Under Assumption 1, the supremum in the above condition is zero at a unique point $\nu(s^a)$ (Lattimore and Szepesvári [2020]), hence $\nu(s^a)$ is the root of the function

$$\nu \mapsto \sup_{\tau \in \mathcal{T}^a} \mathbb{E}_{s^a} \left[\sum_{t=0}^{\tau-1} \gamma^t (R(S_t^a) - \nu) \right]. \quad (5)$$

3 Preliminaries

This section extends the results on the optimal stopping value approximation introduced in Chen and Goldberg [2018] to reward processes that are not restricted to be non-negative, which are needed to develop our results. Section 4 translates these results to the setting of the family of alternative bandit processes.

3.1 Optimal stopping approximation

When considering the behavior of only one of the A arms, the superscript a is dropped from the state, filtration, and set of stopping times. Let $\llbracket 0, t \rrbracket = \{0, 1, \dots, t\}$, $\mathcal{S}_{\llbracket 0, t \rrbracket}$ contain the realisations of \mathcal{S} up to time t , and let \mathcal{F}_t be the smallest sigma algebra for which $\mathcal{S}_{\llbracket 0, t \rrbracket}$ is measurable. For $N \in \mathbb{N}$, let \mathcal{T}_N denote the set of integer-valued stopping times τ adapted to $(\mathcal{F}_t)_{t \in [N]}$ such that $\tau \in [N]$ almost surely. We assume that \mathcal{S} is a Polish space, ensuring the existence of regular conditional probabilities for $(\mathcal{S}_t)_t$ (e.g., Athreya and Lahiri [2006, Theorem 12.3.1]). Let $Z_t = g_t(\mathcal{S}_{\llbracket 0, t \rrbracket})$ for measurable real-valued functions $(g_t)_{t \in [N]}$. The random variable Z_t is assumed integrable (on the probability space for \mathcal{S}) for all t . The goal is to compute

$$\inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau]. \quad (6)$$

The following two results extend Theorems 1 and 2 in Chen and Goldberg [2018] to remove the restriction that Z is non-negative. The proofs readily follow along the lines of the proofs in Chen and Goldberg [2018]. Theorem 1 expresses (6) as an infinite sum. For Z_t bounded, Theorem 2 provides an error bound for truncation of the infinite sum.

Theorem 1 (OPTIMAL STOPPING VALUE REPRESENTATION).

$$\inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau] = \sum_{k=1}^{\infty} \mathbb{E} \left[\min_{u \in [N]} Z_u^{(k)} \right],$$

where for all $k \in \mathbb{N}$ and $t \in [N]$

$$Z_t^{(1)} = Z_t, \quad (7)$$

$$Z_t^{(k+1)} = Z_t^{(k)} - \mathbb{E} \left[\min_{u \in [N]} Z_u^{(k)} \middle| \mathcal{F}_t \right]. \quad (8)$$

Proof. The proof follows the proof of Lemma 1 and Theorem 1 in Chen and Goldberg [2018] noting that, even though Z is not assumed non-negative in this case, the sequence $(Z_t^{(k)})_{k \geq 2}$ remains a non-negative decreasing sequence of random variables for all $t \in [N]$. Hence $\lim_{k \rightarrow \infty} \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau^{(k)}] = 0$ still holds. \square

Theorem 2 (OPTIMAL STOPPING APPROXIMATION). *Suppose $Z_t \in [a, b]$ almost surely for all $t \in [N]$ for some $a, b \in \mathbb{R}$ such that $a < b$, then*

$$0 \leq \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau] - \sum_{k=1}^K \mathbb{E} \left[\min_{u \in [N]} Z_u^{(k)} \right] \leq \frac{b-a}{K+1}. \quad (9)$$

Proof. The proof for $a = 0, b = 1$ follows from Chen and Goldberg [2018, Theorem 2]. The extension to general closed intervals follows by considering the following mapping between stochastic processes

$$T(Z) = \left(Z_t - \mathbb{E} \left[\min_{u \in [N]} Z_u \middle| \mathcal{F}_t \right] \right)_{t \in \mathbb{N}}.$$

We have that $T((c_1 - c_2)Z + c_3) = (c_1 - c_2)T(Z)$ for all $c_1, c_2, c_3 \in \mathbb{R}$. Hence, letting $\tilde{Z} = (Z - a)/(b - a)$ we have from Chen and Goldberg [2018, Theorem 2] that

$$\inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau^{(k+1)}] = \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[T^k(Z)_\tau] = (b-a) \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[T^k(\tilde{Z})_\tau] = (b-a) \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[\tilde{Z}_\tau^{(k+1)}] \leq \frac{b-a}{K+1}.$$

\square

3.2 Simulation approximation

Following Chen and Goldberg [2018], the sum of expectations in (9) may be approximated via simulation. Let $[v, w]$ denote concatenation of the vectors v and w . Let $\mathbf{1}_m$ be the all-ones vector in \mathbb{R}^m for all $m \in \mathbb{N}$. For each index $i \in \cup_{k=1}^K \mathbb{N}^k$ let \mathbf{S}^i be versions of \mathbf{S} such that $\mathbf{S}^i = \mathbf{S}^j$ if $j = [i, \mathbf{1}_m]$, and \mathbf{S}^i is independent of \mathbf{S}^j otherwise.

Let $K \in \mathbb{N}$ and $n(i, k) \in \mathbb{N}$ for all i, k . Recall (7) and (8), for all $k < K$ define the random processes

$$Z_{i,t,n}^{(1)} = g_t(\mathbf{S}_{[0,t]}^i), \quad (10)$$

$$Z_{i,t,n}^{(k+1)} = Z_{[i,1],t,n}^{(k)} - \frac{1}{n(i,k)} \sum_{j=2}^{n(i,k)+1} \left(\min_{u \in [N]} Z_{[i,j],u,n}^{(k)} \mid \{\mathbf{S}_{[0,t]}^{[i,j]} = \mathbf{S}_{[0,t]}^i\} \right), \quad (11)$$

where $Z_{[i,j],u,n}^{(k)} \mid \{\mathbf{S}_{[0,t]}^{[i,j]} = \mathbf{S}_{[0,t]}^i\}$ denotes the random variable $Z_{[i,j],u,n}^{(k)}$ conditioned on the event that the paths of the two processes $\mathbf{S}^{[i,j]}$ and \mathbf{S}^i are equal up to time t . After time t they continue independently (as $j \geq 2$). Note that this random variable is well defined as the regular conditional probabilities for $(\mathbf{S}_t)_t$ exist (Section 3.1). The requirement $\mathbf{S}^i = \mathbf{S}^j$ if $j = [i, \mathbf{1}_m]$ induces that the ‘‘right’’ partial paths of \mathbf{S}^i used at a lower level are used to construct values $Z_{i,t,n}^{(k)}$ at higher level k . To illustrate our notation and the relation between (7), (8) and (10), (11), note that $Z_{[2,j_1],t,n}^{(2)}$ equals

$$\begin{aligned} & Z_{[2,j_1,1],t,n}^{(1)} - \frac{1}{n([2,j_1],1)} \sum_{j_2=2}^{n([2,j_1],1)+1} \left(\min_{u \in [N]} Z_{[2,j_1,j_2],u,n}^{(1)} \mid \{\mathbf{S}_{[0,t]}^{[2,j_1,j_2]} = \mathbf{S}_{[0,t]}^{[2,j_1]}\} \right) \\ & \approx Z_{[2,j_1,1],t,n}^{(1)} - \mathbb{E} \left[\min_{u \in [N]} Z_{[2,j_1,1],u,n}^{(1)} \mid \{\mathbf{S}_{[0,t]}^{[2,j_1,1]} = \mathbf{S}_{[0,t]}^{[2,j_1]}\} \right] = Z_{[2,j_1,1],t,n}^{(1)} - \mathbb{E} \left[\min_{u \in [N]} Z_{[2,j_1,1],u,n}^{(1)} \mid \mathcal{F}_t^{[2,j_1,1]} \right], \end{aligned}$$

where $Z_{[2,j_1,j_2],t,n}^{(1)} = g_t(\mathbf{S}_{[0,t]}^{[2,j_1,j_2]})$, $\mathcal{F}_t^{[2,j_1,1]}$ is the sigma algebra generated by $\mathbf{S}_{[0,t]}^{[2,j_1,1]}$, and the approximation is exact in the limit $n([2,j_1],1) \rightarrow \infty$ by the law of large numbers. Note that the process $\mathbf{S}^{[2,j_1,1]}$ used to determine $Z_{[2,j_1,1],t,n}^{(1)}$ equals $\mathbf{S}^{[2,j_1]}$. Hence, corresponding to conditioning on \mathcal{F}_t above, in order to determine $Z_{[2,j_1],t,n}^{(2)}$ it is assumed that the history of the Markov process up to time t is the same for all $Z_{[2,j_1,j_2],t,n}^{(1)}$ after which the processes continue independently.

Consider the following random variable projected on $[a, b]$:

$$V_n^{(K)} = \max \left(a, \min \left(b, \frac{1}{n(K,K)} \sum_{j=1}^{n(K,K)} \sum_{k=1}^K \min_{u \in [N]} Z_{[k,j],u,n}^{(k)} \right) \right). \quad (12)$$

By the law of large numbers and Theorem 1, we have that

$$\lim_{K \rightarrow \infty} \lim_{n \rightarrow \infty} V_n^{(K)} = \lim_{K \rightarrow \infty} \sum_{k=1}^K \mathbb{E} \left[\min_{u \in [N]} Z_u^{(k)} \right] = \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau],$$

where $n \rightarrow \infty$ indicates the element-wise limit $n(i, k) \rightarrow \infty$ for all (i, k) .

Following Chen and Goldberg [2018], we may show that for every $\xi > 0$ and $\delta \in (0, 1)$ we can choose a function $n_{\xi,\delta}$, given implicitly in Algorithms \mathcal{B}^k and $\hat{\mathcal{B}}^k$ in Chen and Goldberg [2018, Pages 27 and 30], such that

$$\mathbb{P} \left(\left| \sum_{k=1}^K \mathbb{E} \left[\min_{u \in [N]} Z_u^{(k)} \right] - V_{n_{\xi,\delta}}^{(K)} \right| \leq \xi/2 \right) \geq 1 - \delta.$$

This statement was shown in Chen and Goldberg [2018] for non-projected $V_{n_{\xi,\delta}}^{(K)}$. However, the projection on $[a, b]$ can only reduce the error $|\sum_{k=1}^K \mathbb{E}[\min_{u \in [N]} Z_u^{(k)}] - V_{n_{\xi,\delta}}^{(K)}|$ so the statement still holds for $V_{n_{\xi,\delta}}^{(K)}$ as defined in (12). Hence, choosing $K(\xi) = \lfloor 2(b-a)/\xi \rfloor$ we have by Theorem 2 and the triangle inequality that

$$\mathbb{P} \left(\left| \inf_{\tau \in \mathcal{T}_N} \mathbb{E}[Z_\tau] - V_{n_{\xi,\delta}}^{(K(\xi))} \right| \leq \xi \right) \geq 1 - \delta.$$

4 Gittins index approximation

This section first introduces our main method, which is a sampling-based method for Gittins index approximation. Section 4.1 develops finite-time bounds for the approximation, and Section 4.2 develops asymptotic convergence results.

Combining the results of Sections 2 and 3.1, we define for some $R_\ell, R_u \in \mathbb{R}$ such that $R_\ell < R_u$

$$\tilde{Z}_t(\nu) = g_t^\nu(\mathbf{S}_{[0,t]}) = \frac{1-\gamma}{2(R_u - R_\ell)(1-\gamma^N)} \sum_{u=0}^{t-1} \gamma^u (\nu - R(\mathbf{S}_u)), \quad (13)$$

which is the argument in (5), scaled by $c = \frac{1-\gamma}{2(R_u - R_\ell)(1-\gamma^N)}$ for later convenience.

We now introduce our sampling-based Gittins index approximation.

SAMPLING-BASED GITTINS INDEX APPROXIMATION (SBGIA)

- **Approximation I: truncation**

Truncate the support of \tilde{Z}_t and the time horizon for the optimal stopping problem (4):

$$\nu_\sigma(\mathbf{s}) = \sup \left\{ \nu : \inf_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_\mathbf{s}[\tilde{Z}_\tau(\nu)] \leq 0 \right\} \approx \sup \left\{ \nu : \sup_{\tau \in \mathcal{T}} \mathbb{E}_\mathbf{s} \left[\sum_{t=0}^{\tau-1} \gamma^t (R(\mathbf{S}_t) - \nu) \right] \geq 0 \right\}, \quad (14)$$

where the infimum is taken over the set of stopping times $\mathcal{T}_\sigma = \{\tau \in \mathcal{T} : \tau \leq \sigma\}$ with, for a choice of $N \in \mathbb{N}$,

$$\sigma = \sigma_H \wedge N, \quad \sigma_H = \inf\{t \in \mathbb{N} : R(\mathbf{S}_t) \notin [R_\ell, R_u]\} \quad (15)$$

where the minimum operator is denoted with \wedge . Note that ν_σ only considers $R(\mathbf{S}_t) \in [R_\ell, R_u]$. Hence, using the stopped (bounded) process $Z = \tilde{Z}^\sigma$ such that $\tilde{Z}_t^\sigma = Z_{t \wedge \sigma}$ for all t , ν_σ in (14) can also be formulated as

$$\nu_\sigma(\mathbf{s}) = \sup \left\{ \nu : \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_\mathbf{s}[Z_\tau(\nu)] \leq 0 \right\}. \quad (16)$$

Boundedness of Z allows using the results stated in Section 3.1.

- **Approximation II: simulation**

Highlighting the dependence on the state \mathbf{s} and current estimate ν only, we sample $V_\mathbf{s}(\nu) = V_{\mathbf{s}, n_{\xi, \delta}}^{(K(\xi), N)}(\nu)$ truncated to $[-1/2, 1/2]$ by sampling the respective processes \mathbf{S}^i , from the Markov kernel $\mathbb{P}_\mathbf{s}$ starting from state \mathbf{s} , needed to determine $Z_{i,t,n_{\xi, \delta}}^{(K(\xi))}$ in (10), (11) and combining them in (12) such that

$$\mathbb{P} \left(\left| \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_\mathbf{s}[Z_\tau(\nu)] - V_\mathbf{s}(\nu) \right| \leq \xi \right) \geq 1 - \delta. \quad (17)$$

Using this sampling procedure, we approximate $\nu_\sigma(\mathbf{s})$ using stochastic approximation (Borkar [2008]), i.e., a stochastic root-finding procedure, starting from an initial point $\nu_0(\mathbf{s}) \in [R_\ell, R_u]$ such that

$$\nu_{m+1}(\mathbf{s}) = \nu_m(\mathbf{s}) - \alpha_m V_{\mathbf{s}, m}(\nu_m(\mathbf{s})), \quad (18)$$

where $V_{\mathbf{s}, m}$ are independent versions of $V_\mathbf{s}$, and $(\alpha_m)_m$ is a possibly stochastic, predictable non-negative sequence of step-sizes in \mathbb{R} . We collect $\nu_M(\mathbf{s})$ as our sampling-based approximation of the Gittins index $\nu(\mathbf{s})$, where M is defined according to a certain (user-defined) stopping criterion (see Remark 2).

4.1 Finite-time error bounds

In this section, we first derive a truncation error bound for the first-stage approximation (Theorem 3). Subsequently, we couple the stochastic approximation iterates from the second-stage approximation to stochastic approximation iterates from a continuous increasing function (Lemma 1). Using a mean-squared error recursion result for these coupled sequences (Theorem 4), we then construct a confidence interval for the Gittins index in finite-time, where ‘‘finite-time’’ pertains to the number of stochastic approximation iterates (Theorem 5). Using finite-time bounds, we construct an ϵ -optimal policy for the family of alternative bandit processes (Theorem 6).

An upper bound on the error when approximating $\nu(\mathbf{s})$ by the truncation-based index $\nu_\sigma(\mathbf{s})$ as defined in (14) is given in the following theorem, which holds for a general stopping time σ . A similar bound was given in Wang [1997] for a fixed truncation N of the time horizon.

Theorem 3 (TRUNCATION ERROR BOUND). *Under Assumption 1 there is a unique real number $\nu_\sigma(\mathbf{s})$ attaining the supremum*

$$\sup \left\{ \nu : \inf_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_{\mathbf{s}}[\tilde{Z}_\tau(\nu)] \leq 0 \right\}.$$

In fact, $\nu_\sigma(\mathbf{s})$ is the unique root of

$$f_{\mathbf{s}}^\sigma : \nu \mapsto \inf_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_{\mathbf{s}}[\tilde{Z}_\tau(\nu)].$$

The infimum in the condition is attained for some $\tau_\sigma(\nu, \mathbf{s}) \in \mathcal{T}_\sigma$.

Furthermore, with $x^+ = \max(x, 0)$ for $x \in \mathbb{R}$, it holds that

$$0 \leq \nu(\mathbf{s}) - \nu_\sigma(\mathbf{s}) \leq \frac{\mathbb{E}_{\mathbf{s}}[\gamma^\sigma \nu(\mathbf{S}_\sigma)^+]}{1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma]}.$$

Proof. The complete proof can be found in Appendix A and is outlined here. The first two statements follow by showing that $\nu_\sigma(\mathbf{s})$ equals the Gittins index for the Markov process $\tilde{\mathbf{S}}$ documenting the full history of \mathbf{S} up to and including each time t with rewards $\tilde{R}(\tilde{\mathbf{S}}_t) = R(\mathbf{S}_t)\mathbb{I}(t < \sigma)$, after which we can apply commonly known results for the Gittins index from Lattimore and Szepesvári [2020]. Only the upper bound $\nu(\mathbf{s}) - \nu_\sigma(\mathbf{s}) \leq \mathbb{E}_{\mathbf{s}}[\gamma^\sigma \nu(\mathbf{S}_\sigma)^+] / (1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma])$ is non-trivial and mainly follows from bounding the difference between two optimal stopping values by the optimal stopping value of the difference and using the strong Markov property. \square

We now give an error bound for the sampling-based approximation $\nu_M(\mathbf{s})$ as defined in (18). Note that $Z_t(\nu) \in [-1/2, 1/2]$ almost surely. Hence, according to Theorem 2 we have

$$0 \leq \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_{\mathbf{s}}[Z_\tau(\nu)] - \sum_{k=1}^K \mathbb{E}_{\mathbf{s}} \left[\min_{u \in [N]} Z_u^{(k)}(\nu) \right] \leq \frac{1}{K+1}.$$

Defining the functions

$$\tilde{f}_{\mathbf{s}} : \nu \mapsto \mathbb{E}_{\mathbf{s}}[V_{\mathbf{s}}(\nu)], \quad f_{\mathbf{s}} : \nu \mapsto \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_{\mathbf{s}}[Z_\tau(\nu)], \quad (19)$$

we have by (17) and by Jensen's inequality

$$|\tilde{f}_{\mathbf{s}}(\nu) - f_{\mathbf{s}}(\nu)| \leq \mathbb{E}_{\mathbf{s}} \left| \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_{\mathbf{s}}[Z_\tau(\nu)] - V_{\mathbf{s}}(\nu) \right| \leq \delta + \xi =: B(\delta, \xi). \quad (20)$$

Using these results, we derive mean-squared error bounds for our approximation method. We do this by defining coupled stochastic approximation iterates based on the function $f_{\mathbf{s}}$ that lie almost surely below and above the sequence $(\nu_m(\mathbf{s}))_m$ defined in (18). As we do not have access to $f_{\mathbf{s}}$ or even unbiased estimates of $f_{\mathbf{s}}$, these iterates cannot directly be simulated, but we can bound their distance to $(\nu_m(\mathbf{s}))_m$, and show finite-time convergence results for these sequences, which can then be used to derive finite-time convergence bounds for $(\nu_m(\mathbf{s}))_m$. To this end, let

$$\epsilon_m = V_{\mathbf{s},m}(\nu_m(\mathbf{s})) - \tilde{f}_{\mathbf{s}}(\nu_m(\mathbf{s})), \quad (21)$$

which is a martingale difference sequence with respect to the natural filtration $(\mathcal{F}_m^\epsilon)_m$ w.r.t. $(\epsilon_m)_m$ as $V_{\mathbf{s},m}$ (hence $\tilde{f}_{\mathbf{s}}$) is bounded and

$$\mathbb{E}[\epsilon_m \mid \mathcal{F}_{m-1}^\epsilon] = \mathbb{E}[V_{\mathbf{s},m}(\nu_m(\mathbf{s})) \mid \nu_m(\mathbf{s})] - \tilde{f}_{\mathbf{s}}(\nu_m(\mathbf{s})) = 0,$$

since $\nu_m(\mathbf{s})$ is a function of $(\epsilon_{m'})_{m'=1}^{m-1}$.

Let

$$\bar{\nu}_0(\mathbf{s}) = \nu_0(\mathbf{s}), \quad \bar{\nu}_{m+1}(\mathbf{s}) = \bar{\nu}_m(\mathbf{s}) - \alpha_m (f_{\mathbf{s}}(\bar{\nu}_m(\mathbf{s})) - B(\delta, \xi) + \epsilon_m), \quad (22)$$

$$\underline{\nu}_0(\mathbf{s}) = \nu_0(\mathbf{s}), \quad \underline{\nu}_{m+1}(\mathbf{s}) = \underline{\nu}_m(\mathbf{s}) - \alpha_m (f_{\mathbf{s}}(\underline{\nu}_m(\mathbf{s})) + B(\delta, \xi) + \epsilon_m), \quad (23)$$

where $(\alpha_m)_m$ is the same sequence as in (18).

Lemma 1. Assume the step-size sequence $(\alpha_m)_m$ is such that $\sup_m \alpha_m \leq 2(R_u - R_\ell)$ almost surely. For $m = 0, 1, 2, \dots$, we have

$$\underline{\nu}_m \leq \nu_m \leq \bar{\nu}_m. \quad (24)$$

Proof. Observe that $\bar{\nu}_0 = \underline{\nu}_0 = \nu_0$, so that (24) is satisfied for $m = 0$. Now assume $\underline{\nu}_m \leq \nu_m \leq \bar{\nu}_m$ for some $m \geq 0$. We have

$$\begin{aligned} \bar{\nu}_{m+1}(\mathbf{s}) - \nu_{m+1}(\mathbf{s}) &= \bar{\nu}_m(\mathbf{s}) - \nu_m(\mathbf{s}) - \alpha_m \left(f_{\mathbf{s}}(\bar{\nu}_m(\mathbf{s})) - (\tilde{f}_{\mathbf{s}}(\nu_m(\mathbf{s})) + B(\delta, \xi)) \right) \\ &\geq \bar{\nu}_m(\mathbf{s}) - \nu_m(\mathbf{s}) - \alpha_m (f_{\mathbf{s}}(\bar{\nu}_m(\mathbf{s})) - f_{\mathbf{s}}(\nu_m(\mathbf{s}))) \\ &\geq \left(1 - \frac{\alpha_m}{2(R_u - R_\ell)} \right) (\bar{\nu}_m(\mathbf{s}) - \nu_m(\mathbf{s})) \geq 0, \end{aligned}$$

where the second statement follows from (20), and the last statement follows as for any $\nu_1, \nu_2 \in \mathbb{R}$

$$\frac{\min \left(\frac{1-\gamma}{1-\gamma^N}(\nu_1 - \nu_2), \nu_1 - \nu_2 \right)}{2(R_u - R_\ell)} \leq f_{\mathbf{s}}(\nu_1) - f_{\mathbf{s}}(\nu_2) \leq \frac{\max \left(\frac{1-\gamma}{1-\gamma^N}(\nu_1 - \nu_2), \nu_1 - \nu_2 \right)}{2(R_u - R_\ell)}, \quad (25)$$

which can easily be shown from the definition of $f_{\mathbf{s}}$, using the fact that the suprema are attained at unique stopping times.

Similarly, if $\underline{\nu}_m(\mathbf{s}) \leq \nu_m(\mathbf{s})$ for some m , then $\underline{\nu}_{m+1}(\mathbf{s}) \leq \nu_{m+1}(\mathbf{s})$. The proof is completed by induction. \square

The next theorem gives the mean-squared error between either sequence $\bar{\nu}_m(\mathbf{s})$, $\underline{\nu}_m(\mathbf{s})$ and a limit point.

Theorem 4 (MEAN-SQUARED ERROR RECURSION FOR COUPLED SEQUENCES). For $\omega^* \in \mathbb{R}$, initial point $\omega_0 \in \mathbb{R}$, and martingale difference sequence (21), let the sequence $(\omega_m)_m$ be defined as

$$\omega_{m+1} = \omega_m - \alpha_m (f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*) + \epsilon_m).$$

Assume the step-size sequence is such that $\sup_m \alpha_m \leq 2(R_u - R_\ell)$ almost surely. Then for all m

$$\mathbb{E}[(\omega^* - \omega_{m+1})^2] \leq \mathbb{E}[(1 - c\alpha_m)^2 (\omega^* - \omega_m)^2] + \mathbb{E}[\alpha_m^2] \mathbb{E}[\epsilon_m^2].$$

Proof. The full proof can be found Appendix A. The result is obtained by using conditional independence of the martingale difference sequence ϵ_m , rewriting the difference between $f_{\mathbf{s}}(\omega^*)$ and $f_{\mathbf{s}}(\omega_m)$ to a scaled difference between ω^* and ω_m using (25) and the law of total expectation. \square

Observe that by the assumption on α_m we have $c\alpha_m < 1$ (for $N > 1$) so that the influence of the initial difference between ω_0 and ω^* decays exponentially. The squared difference of the iterate ω_m and ω^* converges to a value that depends on the second moment of the martingale differences and the step-size sequence $(\alpha_m)_m$.

Different choices of step-size sequences yield different upper bounds on the rates of convergence. We examine two standard choices.

Example 1 (Constant step-size). Let $\alpha_m = \alpha$ for all m with $\alpha \leq 2(R_u - R_\ell)$. Let $B_0^2 = (\omega_0 - \omega^*)^2$, and assume $\max_m \mathbb{E}[\epsilon_m^2] \leq v_\epsilon^2$. From the recursion for ω_m , we obtain that

$$\mathbb{E}[(\omega^* - \omega_m)^2] \leq B_0^2 (1 - c\alpha)^{2m} + \frac{v_\epsilon^2 \alpha}{c}.$$

Example 2 (Linear step-size). Let $\alpha_m = A/m$ for all m , with $A > 1/c$. Assume $\max_m \mathbb{E}[\epsilon_m^2] \leq v_\epsilon^2$. Then, by Theorem 4 we have for all $m_0 \geq A/(2(R_u - R_\ell))$ that

$$\mathbb{E}[(\omega^* - \omega_{m+1})^2] \leq (1 - cA/m)^2 \mathbb{E}[(\omega^* - \omega_m)^2] + A^2 v_\epsilon^2 / m^2 \leq (1 - cA/m) \mathbb{E}[(\omega^* - \omega_m)^2] + A^2 v_\epsilon^2 / m^2.$$

Hence, by Chung [1954, Lemma 1] we have for all $m \geq A/(2(R_u - R_\ell))$ that

$$\mathbb{E}[(\omega^* - \omega_m)^2] \leq \frac{A^2 v_\epsilon^2}{m(cA - 1)} + o(m^{-2} + m^{-cA}).$$

In the above two examples, we see that we can choose the step-size sequence and stopping point m such that $\mathbb{E}[(\omega^* - \omega_m)^2]$ is arbitrarily small. We can construct confidence intervals for ω^* using Chebyshev's inequality:

$$\mathbb{P}(|\omega^* - \omega_m| > \xi) \leq \frac{\mathbb{E}[(\omega^* - \omega_m)^2]}{\xi^2} \leq \delta.$$

Observe that f_s is a strictly increasing surjective function on \mathbb{R} . We may therefore

$$\text{choose } \bar{\nu}(s), \underline{\nu}(s) \text{ such that } f_s(\bar{\nu}(s)) = B(\delta, \xi), \quad f_s(\underline{\nu}(s)) = -B(\delta, \xi). \quad (26)$$

We may now construct a finite-time confidence interval for the Gittins index $\nu(s)$ for finite m , which can be made arbitrary small by choosing $\mathbb{E}_s[\sigma]$, m large and ξ, δ small.

Theorem 5 (FINITE-TIME CONFIDENCE INTERVAL). *Let $\bar{\nu}(s), \underline{\nu}(s)$ be chosen according to (26). Then*

$$\bar{\nu}(s) - \underline{\nu}(s) \leq 2B(\delta, \xi)/c, \quad (27)$$

which can be made arbitrarily small by suitable choice of ξ and δ .

Let $(\alpha_m)_m$ and M be chosen such that $\sup_m \alpha_m \leq 2(R_u - R_\ell)$, and

$$\mathbb{P}(|\bar{\nu}(s) - \bar{\nu}_M(s)| \leq \xi_2) \geq 1 - \delta_2/2, \quad \mathbb{P}(|\underline{\nu}(s) - \underline{\nu}_M(s)| \leq \xi_2) \geq 1 - \delta_2/2. \quad (28)$$

Then, with probability at least $1 - \delta_2$

$$\nu(s) \in \left(\nu_M(s) - \xi_2 - 2B(\delta, \xi)/c, \nu_M(s) + \xi_2 + 2B(\delta, \xi)/c + \frac{\mathbb{E}_s[\gamma^\sigma \nu(\mathbf{S}_\sigma)^+]}{1 - \mathbb{E}_s[\gamma^\sigma]} \right). \quad (29)$$

Proof. From (25) we obtain that

$$2B(\delta, \xi) = f_s(\bar{\nu}(s)) - f_s(\underline{\nu}(s)) \geq c(\bar{\nu}(s) - \underline{\nu}(s)),$$

which implies (27).

Observe that by (26), $\nu_\sigma(s) \in [\underline{\nu}(s), \bar{\nu}(s)]$ as f_s is increasing and $f_s(\nu_\sigma(s)) = 0$. Hence, by a union bound for the complements of the events in (28), and inequality (27), we have with probability larger than $1 - \delta_2$ that the following confidence interval holds for the second-stage approximation

$$\nu_\sigma(s) \in (\nu_m(s) - \xi_2 - 2B(\delta, \xi)/c, \nu_m(s) + \xi_2 + 2B(\delta, \xi)/c).$$

The result follows from Theorem 3. \square

The confidence interval (29) can be used to construct an ϵ -optimal policy for any FABP. For this we first need to show that the values of $(\nu_m)_m$ are restricted to a closed bounded interval. The proof of this lemma can be found in Appendix A.

Lemma 2. *Let $\sup_m \alpha_m \leq M_\alpha < \infty$ almost surely. We have almost surely that for all m*

$$\nu_m(s) \in [R_\ell - M_\alpha/2, R_u + M_\alpha/2].$$

For the family of alternative bandit processes, we can now define an ϵ -optimal policy, which we will denote the SBGIA policy (SBGIAP).

Theorem 6 (ϵ -OPTIMAL POLICY FOR FABP). *Let $\sup_m \alpha_m \leq 2(R_u - R_\ell) < \infty$ almost surely. Let π be the randomized Markov policy such that for all histories $h \in \mathcal{H}$*

$$\pi(h) = \arg \max_{a \in [A]} \nu_M(\mathbf{s}_h^a),$$

where $(\nu_m)_m$ is determined by (14) – (18), and M, σ, δ, ξ are chosen such that for some $\epsilon > 0$

$$\mathbb{P}(|\nu(\mathbf{s}_h^a) - \nu_M(\mathbf{s}_h^a)| \leq (1 - \gamma)^2 \epsilon / 4) \geq 1 - (1 - \gamma)^2 \epsilon / (4AD(\mathbf{s}_h^a)) \quad \forall a \in [A], \quad (30)$$

for $D(\mathbf{s}_h^a) = (1 - \gamma)C(\mathbf{s}_h^a) + \max(|R_\ell|, |R_u|) + M_\alpha/2$. Then, the policy π is ϵ -optimal for the family of alternative bandit processes, i.e.,

$$\mathbb{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{S}_{N_{A_t, t}}^{A_t}) \right] - \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{S}_{N_{A_t, t}}^{A_t}) \right] \leq \epsilon.$$

Proof. Note that (30) is possible due to (29). in Glazebrook [1982], it is shown that for any stationary policy π

$$\mathbb{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t R \left(\mathbf{S}_{N_{A_t,t}}^{A_t} \right) \right] - \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R \left(\mathbf{S}_{N_{A_t,t}}^{A_t} \right) \right] \leq \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(\max_{a \in [A]} \nu(\mathbf{S}_{N_{a,t}}^a) - \nu \left(\mathbf{S}_{N_{A_t,t}}^{A_t} \right) \right) \right] / (1 - \gamma). \quad (31)$$

By our choice of m, σ, δ, ξ we have by a union bound over a for the events in (30) that for any time point t

$$\mathbb{P} \left(\max_a |\nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^a)| > (1 - \gamma)^2 \epsilon / 4 \right) \leq (1 - \gamma)^2 \epsilon / (4D(\mathbf{S}_t^a)).$$

Combining Lemma 2 with the fact that $\nu(\mathbf{S}_t^a) \leq (1 - \gamma)C(\mathbf{S}_t^a)$ yields $|\nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^a)| \leq D(\mathbf{S}_t^a)$ for all m , hence

$$\mathbb{E} \left[\max_a |\nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^a)| \right] \leq (1 - \gamma)^2 \epsilon / 4 + (1 - \gamma)^2 \epsilon / 4 = (1 - \gamma)^2 \epsilon / 2.$$

Using this in the right-hand side of (31) scaled by $(1 - \gamma)$ gives

$$\begin{aligned} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(\max_{a \in [A]} \nu(\mathbf{S}_t^a) - \nu \left(\mathbf{S}_t^{A_t} \right) \right) \right] &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(\max_{a \in [A]} \nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^{A_t}) + \nu_m(\mathbf{S}_t^{A_t}) - \nu \left(\mathbf{S}_t^{A_t} \right) \right) \right] \\ &\leq \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(\max_{a \in [A]} (\nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^a)) + \nu_m(\mathbf{S}_t^{A_t}) - \nu \left(\mathbf{S}_t^{A_t} \right) \right) \right] \\ &\leq \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi} \left[2 \max_{a \in [A]} |\nu(\mathbf{S}_t^a) - \nu_m(\mathbf{S}_t^a)| \right] \leq (1 - \gamma) \epsilon. \end{aligned}$$

□

4.2 Asymptotic convergence results

This section investigates convergence properties of the stochastic process defined in (18). In the previous section, we have shown that after a finite, known, amount of iterations, the iterates $\nu_m(\mathbf{s})$ lie in an interval containing $\nu(\mathbf{s})$ with high probability. The length of this interval depends on the choice of K and n . In this section, we let K and n be constants, not depending on ξ, δ , and investigate convergence and asymptotic normality of the iterates $\nu_m(\mathbf{s})$ when m goes to infinity, under different choices of the step-size sequence. First, we show that when the step-size sequence almost surely satisfies the Robbins-Monro conditions (Robbins and Monro [1951]), the stochastic approximation iterates converge almost surely (Theorem 7). Then, we show that if we instead take a constant step-size sequence, the iterates converge in mean-square to the set of roots (Theorem 8). Lastly, under stronger conditions, we show that we can construct an adaptive stochastic approximation procedure (Lai and Robbins [1979]), where a central limit theorem holds for the stochastic approximation iterates (Theorem 9). In the following, we let $V_s(\nu) = V_{s,n}^{(K,N)}(\nu)$ for fixed choices of K, N, n , and define \tilde{f}_s and ϵ_m as in (19), (21), respectively.

The sequence (18) relates to an Euler scheme for the first-order scalar autonomous ordinary differential equation (ODE) (Borkar [2008]),

$$\frac{d}{dx} \nu(x) = -\tilde{f}_s(\nu(x)). \quad (32)$$

This differential equation has as equilibrium points the roots \mathcal{R}_s of the function \tilde{f}_s , provided that a root exists. We show that iteration scheme (18) satisfies the conditions stated in Borkar [2008, Chapter 2], hence almost surely the limiting behavior of the sample paths of the stochastic process (18) equals that of the solution to the above ODE. We then show that the ODE (32) always converges to an equilibrium point, irrespective of the starting point. It follows that the sample paths of (18) almost surely converge to a random variable $\tilde{\nu}(\mathbf{s}) \in \mathcal{R}_s$ such that $\tilde{f}_s(\tilde{\nu}(\mathbf{s})) = 0$.

To make this formal, we define an *internally chain transitive invariant set* corresponding to an ODE in accordance with the definition given in Borkar [2008].

Definition 1 (Internally chain transitive invariant set (Borkar [2008])). *A closed set $E \subset \mathbb{R}$ is said to be an internally chain transitive invariant set for the ODE (32) if*

- any trajectory ν of (32) with $\nu(0) \in E$ satisfies $\nu(x) \in E \forall x \in \mathbb{R}$,
- for any $\nu, \nu' \in E$ and any $\epsilon > 0, T > 0$, there exist $n \geq 1$ and points $\nu_0 = \nu, \nu_1, \dots, \nu_{n-1}, \nu_n = \nu'$ in E such that the trajectory of (32) initiated at ν_i meets with the ϵ -neighbourhood of ν_{i+1} for $0 \leq i < n$ after a time $\geq T$.

We first show that \tilde{f}_s is Lipschitz continuous.

Lemma 3. *The function \tilde{f}_s is Lipschitz continuous in ν .*

Proof. We first prove by induction in k that $Z_{i,u,n}^{(k)}$ is Lipschitz continuous for all k, i, u, n , starting with $k = 1$. From (10) and (13) it follows that $Z_{i,u,n}^{(1)}$ is Lipschitz continuous with constant $L = 1/(2(R_u - R_\ell))$ for all i, u, n . Now assume that $Z_{i,u,n}^{(k)}$ is Lipschitz continuous up to some k for all i, u, n . As the minimum, average, and sum of a finite set of Lipschitz continuous functions are Lipschitz continuous, we have by (11) that $Z_{i,u,n}^{(k+1)}$ is also Lipschitz continuous for all possible i, u, n . By induction we have that $Z_{i,u,n}^{(k)}$ is Lipschitz continuous for all k, i, u, n . Now, by (12)

$$\tilde{f}_s(\nu) = \mathbb{E}_s \left[\max \left(-1/2, \min \left(1/2, \frac{1}{n(K, K)} \sum_{j=1}^{n(K, K)} \sum_{k=1}^K \min_{u \in [N]} Z_{[k, j], u, n}^{(k)}(\nu) \right) \right) \right].$$

Hence, as the expectation, maximum, minimum, and sum of Lipschitz continuous functions are Lipschitz continuous, we have that \tilde{f}_s is Lipschitz continuous. \square

Using this lemma, the next result follows.

Theorem 7 (ALMOST SURE CONVERGENCE OF STOCHASTIC APPROXIMATION ITERATES).

Assume $(\alpha_m)_m$ almost surely satisfies the Robbins-Monro conditions (Borkar [2008]):

$$\sum_{m=1}^{\infty} \alpha_m = \infty, \quad \sum_{m=1}^{\infty} \alpha_m^2 < \infty. \quad (33)$$

The sequence $(\nu_m(s))_m$ generated by (18) almost surely converges to a (possibly sample path dependent) compact connected internally chain transitive invariant set of (32).

Proof. We verify that assumptions (A1 - A4) in Borkar [2008, Chapter 2] are satisfied, from which the result follows by Borkar [2008, Chapter 2, Theorem 2]:

- (A1) We have from Lemma 3 that \tilde{f}_s is Lipschitz continuous.
- (A2) The Robbins-Monro conditions hold by assumption.
- (A3) The sequence $(\epsilon_m)_m$ defined by (21) is a bounded martingale difference sequence.
- (A4) Note that condition (33) is sufficient for the condition on the step-size sequence in Lemma 2, hence we have $\sup_m |\nu_m(s)| < \infty$ almost surely. \square

The next corollary follows.

Corollary 1. *Under Assumption (33), the sequence $(\nu_m(s))_m$ generated by (18) converges almost surely to a random variable $\tilde{\nu}(s) \in [R_\ell, R_u]$ such that*

$$\tilde{f}_s(\tilde{\nu}(s)) = 0.$$

Proof. From the proof of Lemma 2, we have $f_s(\nu) < 0$ for $\nu < R_\ell$ and $f_s(\nu) > 0$ for $\nu > R_u$. Hence, $\frac{d}{dx} \nu(x, s) > 0$ if $\nu < R_\ell$ and $\frac{d}{dx} \nu(x, s) < 0$ if $\nu > R_u$. From Lipschitz continuity of f_s it then follows that no solution of (32) goes to infinity. Hence, by classification of solutions to a first-order scalar autonomous ODE, we know that each solution of (32) must converge to a (semi-)stable point contained in the set \mathcal{R}_s of roots of f_s , which is non-empty and contained in $[R_\ell, R_u]$ by the above discussion. \square

If the step-size sequence is constant, i.e., $\alpha_m \equiv \alpha \in (0, \infty)$ for all m , the following result holds by Borkar [2008, Chapter 9, Theorem 3], which states that $\nu_m(s)$ is close to \mathcal{R}_s in the limit in mean-square, but does not necessarily almost surely converge to a point in \mathcal{R}_s .

Theorem 8 (CONVERGENCE OF STOCHASTIC APPROXIMATION FOR CONSTANT STEP-SIZE).

For a constant $H > 0$

$$\limsup_{m \rightarrow \infty} \mathbb{E} \left[\min_{\nu \in \mathcal{R}_s} (\nu_m(\mathbf{s}) - \nu)^2 \right] \leq H\alpha.$$

Proof. We already saw in the proof of Theorem 7 that assumptions (A1) and (A3) in Borkar [2008, Chapter 2] are satisfied. Furthermore (9.2.1) and (9.2.2) in Borkar [2008] are satisfied as the iterates $(\nu_m(\mathbf{s}))_m$ stay in a closed bounded interval (Lemma 2). \square

Remark 1. In the above results, we have shown two convergence results under two different assumptions on the step-size sequence. Theorem 7 assumes that the Robbins-Monro conditions almost surely hold for the step-size sequence $(\alpha_m)_m$. This is a stronger condition than the one assumed in Section 4.1, where only $\sup_m \alpha_m < M_\alpha$ was assumed for some bound M_α . Under this stronger condition we were able to show that the stochastic process $(\nu_m)_m$ converges to a root of the function \tilde{f}_s and based on a sample path alone, we can determine whether the sequence has converged or not. If we instead take a constant step-size sequence, we could only show that the limit of the stochastic process is close to \mathcal{R}_s , in terms of mean-squared difference. Often the rate of convergence is much higher for constant step-size sequences (Borkar [2008]).

The following lemma gives a recursive formula for the derivative which can be used for selection of the step-size in the adaptive stochastic approximation procedure (18).

Lemma 4. If for all t the cumulative distribution functions of $R(\mathbf{S}_t)$ starting from $\mathbf{S}_0 = \mathbf{s}$ have finitely many jumps, then for all but finitely many points ν the function \tilde{f}_s is differentiable with derivative

$$\frac{d}{d\nu} \tilde{f}_s(\nu) = \sum_{k=1}^K \mathbb{E}_s \left[h_{[k,1],n}^{(k)} \left(U_{[k,1],n}^{(k)}(\nu), \nu \right) \mathbb{I}(\tilde{V}_s(\nu) \in [-1/2, 1/2]) \right],$$

where $\tilde{V}_s(\nu) = \sum_{k=1}^K \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu)$, $U_{\mathbf{i},n}^{(k)}(\nu) = \arg \min_{u \in [N]} Z_{\mathbf{i},u,n}^{(k)}(\nu)$, and

$$\begin{aligned} h_{\mathbf{i},n}^{(1)}(t, \nu) &= (1 - \gamma^t) / (2(R_u - R_\ell)(1 - \gamma^N)), \\ h_{\mathbf{i},n}^{(k)}(t, \nu) &= h_{[\mathbf{i},1],n}^{(k-1)}(t, \nu) - \frac{1}{n(\mathbf{i}, k)} \sum_{j=2}^{n(\mathbf{i}, k)+1} h_{[\mathbf{i},j],n}^{(k-1)}(U_{[\mathbf{i},j],n}^{(k-1)}(\nu), \nu) \{ \mathbf{S}_{[0,t]}^{[\mathbf{i},j]} = \mathbf{S}_{[0,t]}^{\mathbf{i}} \} \quad (k \geq 2). \end{aligned}$$

Proof. The full proof can be found in Appendix A. The result might seem straightforward at first but the situation is more difficult due to the dependence of $U_{\mathbf{i},n}^{(k)}$ on ν . The result is obtained by showing continuity of $U_{\mathbf{i},n}^{(k)}$ for all but at most finitely many points, by giving a coinciding lower and upper bound for the derivative of $\arg \min_{u \in [N]} Z_{\mathbf{i},u,n}^{(k)}$ whenever it exists and then using dominated convergence to show the derivative for the expected sum truncated to $[-1/2, 1/2]$. \square

The next theorem states that using Lemma 4 an adaptive stochastic approximation method can be designed with asymptotically optimal variance (Lai and Robbins [1979]).

Theorem 9 (CENTRAL LIMIT THEOREM FOR STOCHASTIC APPROXIMATION ITERATES).

Let $h_{m,[k,1],n}^{(k)}$, $U_{m,[k,1],n}^{(k)}$ be independent versions (in m) of $h_{[k,1],n}^{(k)}$, $U_{[k,1],n}^{(k)}$, and

$$h_m : \nu \mapsto \sum_{k=1}^K h_{m,[k,1],n}^{(k)}(U_{m,[k,1],n}^{(k)}(\nu), \nu).$$

Let

$$\alpha_m = \frac{1}{|\sum_{\ell=1}^m h_\ell(\nu_\ell(\mathbf{s}))|}.$$

Let \mathcal{V} be the set of points where \tilde{f}_s is differentiable. If $\inf_{\nu \in \mathcal{V}} d/d\nu \tilde{f}_s(\nu) > 0$, then there is a unique point $\tilde{\nu}(\mathbf{s})$ such that $\tilde{f}_s(\tilde{\nu}(\mathbf{s})) = 0$. If the derivative of $\tilde{f}_s^{(K)}$ exists at $\tilde{\nu}(\mathbf{s})$ and $\mathbb{E}[V_s(\tilde{\nu}(\mathbf{s}))^2] > 0$, we have

$$\sqrt{m}(\nu_m(\mathbf{s}) - \tilde{\nu}(\mathbf{s})) \xrightarrow{d} \mathcal{N} \left(0, \frac{\mathbb{E}[V_s(\tilde{\nu}(\mathbf{s}))^2]}{\left(\frac{d}{d\nu} \tilde{f}_s^{(K)}(\tilde{\nu}(\mathbf{s})) \right)^2} \right). \quad (34)$$

Proof. The full proof can be found in Appendix A. We first show that the sequence $m\alpha_m$ converges to a constant, by showing that α_m satisfies the Robbins-Monro conditions almost surely and by applying Corollary 1. After this, we can apply the results in Lai and Robbins [1978], where we have to account for the fact that the residuals are not independent and identically distributed but bounded, and the function is only differentiable at all but finitely many points. \square

Remark 2 (Step-size sequence and stopping criterion). *We propose to use the adaptive step-size sequence $\alpha_m = 1/|\sum_{\ell=1}^m h_\ell(\nu_\ell(\mathbf{s}))|$ for the stochastic approximation sequence (18), and base the stopping criterion for the stochastic approximation sequence on the estimated radius of the confidence interval implied by (55). We stop the stochastic approximation procedure (18) when the estimated confidence radius based on Theorem 9 is small enough, i.e., when*

$$\frac{C_{1-\beta} \sqrt{\frac{1}{m} \sum_{\ell=1}^m (V_{\mathbf{s},\ell}(\nu_\ell(\mathbf{s})))^2}}{\sqrt{m} \frac{1}{m} \sum_{\ell=1}^m h_\ell(\nu_\ell(\mathbf{s}))} \leq \epsilon_\nu, \quad (35)$$

with $C_{1-\beta}$ the level $1 - \beta/2$ quantile of the standard normal distribution, and ϵ_ν a pre-specified tolerance.

Remark 3. *When all values $n(i, k)$ are large enough, we have by Theorem 2*

$$\tilde{f}_{\mathbf{s}}(\nu) \approx \sum_{k=1}^K \mathbb{E}_{\mathbf{s}} \left[\min_{u \in [N]} Z_u^{(k)}(\nu) \right] \leq \inf_{\tau \in \mathcal{T}_N} \mathbb{E}_{\mathbf{s}}[Z_\tau(\nu)] = f_{\mathbf{s}}(\nu), \quad (36)$$

$$f_{\mathbf{s}}(\nu) \leq \sum_{k=1}^K \mathbb{E}_{\mathbf{s}} \left[\min_{u \in [N]} Z_u^{(k)}(\nu) \right] + 1/(K+1) \approx \tilde{f}_{\mathbf{s}}(\nu) + 1/(K+1), \quad (37)$$

where we used $b = -a = 1/2$. As both $\tilde{f}_{\mathbf{s}}$ and $f_{\mathbf{s}}$ are continuous, negative at R_ℓ , positive at C_R , and $f_{\mathbf{s}}$ is increasing, we expect from (36) that the root of $\tilde{f}_{\mathbf{s}}$ is an upper bound of the root of $f_{\mathbf{s}}$. As $\mathbb{E}_{\mathbf{s}}[\min_{u \in [N]} Z_u^{(k)}(\nu)] \geq 0$ for $k \geq 2$, we furthermore expect that the root of $\tilde{f}_{\mathbf{s}}$ decreases to the root of $f_{\mathbf{s}}$, and by (37) we expect that both roots coincide in the limit as $K \rightarrow \infty$. We furthermore have by (25) and (37) that, for large enough values of $n(i, k)$,

$$\tilde{\nu}(\mathbf{s}) - \nu_\sigma(\mathbf{s}) \leq f_{\mathbf{s}}(\tilde{\nu}(\mathbf{s}))/c \leq 1/(c \cdot (K+1)), \quad (38)$$

hence the difference in the two roots is of order $O(1/(K+1))$. In order to get an accurate SBGIA, we hence propose, for a fixed K , to first increase all $n(i, k)$ to large enough values such that the root of $\tilde{f}_{\mathbf{s}}$ has converged. After convergence has occurred, we propose to increase K and to repeat this procedure until the root has also converged in K .

5 Application to Bayesian multi-armed bandits

We introduce the Bayesian multi-armed bandit in Section 5.1 as an application of the FABP. In Section 5.2, we consider outcome distributions from an exponential family. Subsequently, we present results for the SBGIA applied to two Bayesian bandits known from literature, the Bernoulli bandit and Gaussian bandit with known variance. In Section 5.3, we evaluate the performance of the SBGIAP for a novel Gaussian random effects bandit problem.

5.1 Bayesian multi-armed bandit

Consider A distributions with support \mathcal{O} . Selecting distribution a at time $t \in \mathbb{N}$, results in a realisation of the random variable O_t^a that has a known density $p(O_t^a | \theta_a)$ w.r.t. a measure μ , where the unknown parameter θ_a lies in a parameter space Θ (shared for all a). The random variables O_t^a are assumed independent. We perform a Bayesian analysis, where the parameters θ_a are independent a priori and endowed with prior probability measure Π_0^a w.r.t. the Polish space $(\Theta, \mathcal{B}(\Theta))$. Given this probability measure Π_0^a on $(\Theta, \mathcal{B}(\Theta))$, we can determine the predictive distribution

$$p(O | \Pi_0^a) = \int_{\Theta} p(O | \theta) d\Pi_0^a(\theta). \quad (39)$$

A sample from this distribution then, in turn, generates a posterior distribution $\Pi_1^a(\cdot | O, \Pi_0^a)$ by Bayes' rule, i.e.,

$$\Pi_1^a(E | O, \Pi_0^a) = \int_{\Theta} \frac{p(O | \theta)}{\int_{\Theta} p(O | \theta) d\Pi_0^a(\theta)} d\Pi_0^a(\theta) \quad \forall E \in \mathcal{B}(\Theta). \quad (40)$$

We can use the predictive distribution (39) and posterior updating rule (40) to determine A Markov chains (arms), with states $(\Pi_t^a)_{a=1}^A$ corresponding to (posterior) distributions on Θ . The state space of each Markov chain is the space \mathbb{M} of probability measures on $(\Theta, \mathcal{B}(\Theta))$. Furthermore, we endow this state space with the sigma algebra \mathcal{M} , which is the

smallest sigma field making all maps from \mathbb{M} to \mathbb{R} measurable. Then $(\mathbb{M}, \mathcal{M})$ is a Borel space (Ghosal and Van der Vaart [2017, Chapter 3.1]). Each Markov chain $(\Pi_t^a)_{a=1}^A$ has transition kernel

$$\mathbb{P}_{\Pi_t^a}(E) = \int_{\mathcal{O}} \mathbb{I}(\Pi_{t+1}^a(\cdot | O, \Pi_t^a) \in E) p(O | \Pi_t^a) \mu(dO) \quad \forall E \in \mathcal{M}. \quad (41)$$

Our goal is to find a (Markov) policy π to sequentially sample from one of the A arms that maximizes the expected discounted sum of outcomes under the Bayesian model, i.e, to maximize

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t O_{t+1}^{A_t} \right] = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(\Pi_t^{A_t}) \right], \quad (42)$$

where $R(\Pi_t^a) = \int_{\Theta} \mathbb{E}[O | \theta] d\Pi_t^a(\theta)$ is the posterior mean outcome for the current posterior Π_t^a . The equality in (42) follows from Assumption 1 and Fubini's theorem. Letting $\mathcal{S}_t^a = \Pi_t^a$ for all a, t , $\mathcal{S} = \mathbb{M}$, and $\mathcal{G} = \mathcal{M}$, we are in the setting of Section 2 with transition kernel (41) and reward function R .

5.2 Gittins index approximation results

This section considers the FABP with distributions from an exponential family as detailed in Section 5.2.1. Specific results are presented for Bernoulli and Gaussian families in Sections 5.2.2 and 5.2.3.

For determining the SBGIA, we set $\beta = 0.05$ and $\epsilon_\nu = 0.001$ in (35) to estimate a 95% asymptotic confidence interval for $\tilde{v}(s)$ with radius 0.001. We compare the SBGIA with the Calibration method introduced in Gittins [1979], which is a combination of a bisection method and backward induction to obtain the value of the truncated optimal stopping problem in (4). The parameters of the Calibration method are set such that the approximation error is very small, so that we may consider these values to be the true Gittins index values.

5.2.1 Exponential families

Assume the data comes from a distribution belonging to an exponential family, i.e., for known functions ψ, η, ρ, ζ we have

$$p(O^a | \theta_a) = \zeta(O^a) \exp(\eta(\theta_a)^\top \psi(O^a) - \rho(\theta_a)) d\mu(O^a),$$

and

$$p((O_t^a)_t | \theta_a) = \prod_{u=1}^t \zeta(O_u^a) \exp(\eta(\theta_a)^\top \psi(O_u^a) - \rho(\theta_a)) d\mu(O_u^a).$$

Now assume a conjugate prior for this model (Diaconis and Ylvisaker [1979]), with normalizing constant ζ_2

$$p(\theta_a; \Psi_0^a, \kappa_0^a) = \zeta_2(\Psi_0^a, \kappa_0^a) \exp(\eta(\theta_a)^\top \Psi_0^a - \kappa_0^a \rho(\theta_a)).$$

Letting $\Psi_t^a = \Psi_0^a + \sum_{u=1}^t \psi(O_u^a)$ and $\kappa_t^a = \kappa_0^a + t$, we have the following expression for the posterior

$$p(\theta_a | (O_t^a)_t) = \zeta_2(\Psi_t^a, \kappa_t^a) \exp(\eta(\theta_a)^\top \Psi_t^a - \kappa_t^a \rho(\theta_a)) =: p(\theta_a; \Psi_t^a, \kappa_t^a).$$

The vector Ψ_t^a is often referred to as the sufficient statistic, and κ_t^a as the effective number of observations, which is the sum of a prior number of observations κ_0^a and the actual number of observations t for arm a . As the only random element in the posterior is Ψ_t^a , which we will assume to lie in \mathbb{R}^d , the Markov chain $(\Pi_t^a)_t$ can be represented by the time-inhomogeneous Markov chain $(\Psi_t^a, \kappa_t^a)_t$ on the finite-dimensional state space \mathbb{R}^{d+1} with transition dynamics

$$(\Psi_t^a, \kappa_t^a) \rightarrow (\Psi_t^a, \kappa_t^a) + (\psi(O), 1), \quad O \sim p(O | (\Psi_t^a, \kappa_t^a)). \quad (43)$$

The rewards $R(\Psi_t^a, \kappa_t^a) = \int_{\Theta} \mathbb{E}[O | \theta] p(\theta; \Psi_t^a, \kappa_t^a) d\theta$ are the posterior mean outcomes. As our goal is to maximize the expected discounted sum of outcomes under the Bayesian model, letting $\mathcal{S}_t^a = (\Psi_t^a, \kappa_t^a)$ for all a, t , $\mathcal{S} = \mathbb{R}^{d+1}$, and $\mathcal{G} = \mathcal{B}(\mathbb{R}^{d+1})$, we are in the setting of Section 2 with the transition kernel implied by (43) and reward function R .

5.2.2 Bernoulli bandit

We consider the case when O_t^a are Bernoulli distributed with unknown success probability $p_a \in [0, 1]$, i.e.,

$$\mathbb{P}(O^a = 1) = p_a, \quad \mathbb{P}(O^a = 0) = 1 - p_a,$$

which implies that the family of outcome distributions is an exponential family with

$$\zeta \equiv 1, \quad \eta(p_a) = \log(p_a/(1 - p_a)), \quad \Psi(O) = O, \quad \rho(p_a) = -\log(1 - p_a),$$

and μ the counting measure on the nonnegative integers. We assume a conjugate Beta(α_a, β_a) prior on each p_a , hence $\Psi_t^a = \alpha_a + \sum_{u=1}^t O_u^a$ and $\kappa_t^a = \alpha_a + \beta_a + t$. In the following, we drop the superscript a from notation as we consider results for a single arm a only.

Observe that $R(\Psi, \kappa) = \Psi/\kappa$, hence for a fixed horizon $N \in \mathbb{N}$ the Markov reward process starting from (Ψ, κ) is bounded. Thus, in Section 4 we can take $\sigma = N$, and the result of Theorem 3 holds without truncation of the support of the rewards. Using $\nu(\mathcal{S}_N)^+ \leq 1$, and letting the Gittins index approximation after truncation be denoted by $\nu_N(\Psi, \kappa)$ (as $\sigma = N$), we have

$$0 \leq \nu(\Psi, \kappa) - \nu_N(\Psi, \kappa) \leq \frac{\gamma^N}{1 - \gamma^N}, \quad (44)$$

hence for $\epsilon_{\text{trunc}} > 0$

$$N \geq \log_\gamma(\epsilon_{\text{trunc}}/(1 + \epsilon_{\text{trunc}})) \implies \nu(\Psi, \kappa) - \nu_N(\Psi, \kappa) \leq \epsilon_{\text{trunc}}. \quad (45)$$

For the rewards, we have

$$R(\Psi, \kappa + w) \in [\Psi/(\kappa + w), (\Psi + w)/(\kappa + w)] \quad \forall w \in [N].$$

Starting from (Ψ, κ) the rewards lie in the bounded interval $[R_\ell, R_u]$ up to a horizon N for

$$R_\ell = \Psi/(\kappa + N), \text{ and } R_u = (\Psi + N)/(\kappa + N),$$

which can be used in (13).

Table 1 shows the Gittins index values found under the SBGIA and the Calibration method for the Bernoulli bandit with $\gamma = 0.8$. For the SBGIA and the Calibration method, we set $N = 35$, corresponding to a truncation error bound ϵ_{trunc} of 0.0005 according to (45). We set each $n(i, k) = 1$ and considered $K = 1, 2, 3$. For ease of comparison, Table 1 shows $\kappa - \Psi$, the effective number of failures, as the second state variable. Table 1 shows the log computation time (in seconds), estimated bias (SBGIA minus Calibration), root-mean-squared error (RMSE), and average standard deviation over estimates for the SBGIA (calculated as the average deviation in the approximation over states for two independent simulation runs) for each considered value of K . Using (38), the rightmost column of Table 1 shows an estimate of the limit as $K \rightarrow \infty$, found by performing ordinary least squares on the columns $K = 1, 2, 3$ with the line $\nu(s) = a + b/(K + 1)$.

Table 1 shows that the Gittins index is overestimated by the SBGIA for $K = 1$, which is in agreement with our expectation, as the expected minimum is always smaller than the minimum of the expectation over stopping times (see, e.g., Chen and Goldberg [2018]). When K increases, the amount of overestimation decreases, and values of the SBGIA lie above the Gittins index computed by the Calibration method for any state and value of K . It follows from Remark 3 that this behaviour should occur asymptotically when the values of $n(i, k)$ go to infinity; the numerical results show that it also occurs for small values of $n(i, k)$. The estimate of the limit when $K \rightarrow \infty$, shown in the rightmost column of Table 1 has lowest bias and RMSE, indicating a correct assumption on the $O(1/(K + 1))$ convergence rate. The computation time increases more than tenfold with each increase in K , and for $K = 1$ it is already about 100 times larger than the computation time of the Calibration method. The standard deviation over runs (SD) is quite low, around 0.0001 for all values of K , indicating that the estimates are consistent over independent runs.

Table 2 shows the values of the SBGIA and the Calibration method for $K = 2$ and $n(i, 2) = 1, 3, 5$, i.e., only the nested number of simulations for the SBGIA is increased. The CPU time, bias, RMSE, and standard deviations over two independent simulation runs are shown at the bottom of Table 2.

Table 2 shows that when increasing $n(i, k)$ a smaller error (expressed in bias and RMSE) is attained for a lower computational cost in comparison to Table 1. The results in Table 2 hence agree with the proposal in Remark 3, as increasing the number of nested simulations leads to higher precision in less computation time.

Table 1: Gittins index values found under the SBGIA for $K = 1, 2, 3$, the Calibration method, and an estimate for the limit as $K \rightarrow \infty$ for the Bernoulli bandit. We set each $n(i, k) = 1$, $\gamma = 0.8$, $N = 35$, and set a tolerance of $\epsilon_\nu = 0.001$ in the stopping criterion (35) for determining the SBGIA. The computation (CPU) time denotes the time it took to calculate all values in the column and is measured in seconds. The bias, RMSE, and standard deviation (SD) are multiplied by 100, i.e., displayed in percentage points.

General	Ψ	$\kappa - \Psi$	Calibration	$K = 1$	$K = 2$	$K = 3$	Est. limit
	1	1	0.641	0.643	0.642	0.642	0.641
	1	2	0.443	0.447	0.446	0.445	0.442
	1	3	0.332	0.338	0.337	0.335	0.332
	1	4	0.263	0.270	0.268	0.267	0.264
	1	5	0.216	0.224	0.222	0.221	0.218
	1	6	0.183	0.191	0.189	0.188	0.185
	2	1	0.760	0.760	0.760	0.760	0.759
	2	2	0.590	0.592	0.591	0.591	0.590
	2	3	0.476	0.480	0.479	0.478	0.476
	2	4	0.398	0.402	0.401	0.400	0.398
	2	5	0.340	0.345	0.344	0.342	0.340
	2	6	0.296	0.301	0.300	0.299	0.297
	3	1	0.816	0.816	0.816	0.816	0.815
	3	2	0.671	0.673	0.673	0.672	0.671
	3	3	0.566	0.568	0.568	0.567	0.566
	3	4	0.487	0.490	0.489	0.489	0.487
	3	5	0.427	0.430	0.429	0.428	0.427
	3	6	0.379	0.383	0.382	0.381	0.379
	4	1	0.849	0.850	0.849	0.849	0.849
	4	2	0.725	0.726	0.725	0.725	0.724
	4	3	0.628	0.629	0.629	0.629	0.628
	4	4	0.552	0.554	0.554	0.553	0.552
	4	5	0.491	0.494	0.494	0.493	0.492
	4	6	0.443	0.446	0.445	0.444	0.443
	5	1	0.872	0.873	0.872	0.872	0.871
	5	2	0.762	0.763	0.763	0.762	0.762
	5	3	0.674	0.675	0.675	0.674	0.673
	5	4	0.602	0.604	0.603	0.603	0.602
	5	5	0.543	0.545	0.545	0.544	0.543
	5	6	0.494	0.497	0.496	0.496	0.495
	6	1	0.888	0.889	0.888	0.888	0.887
	6	2	0.790	0.791	0.791	0.791	0.790
	6	3	0.709	0.710	0.710	0.709	0.709
	6	4	0.641	0.643	0.642	0.642	0.641
	6	5	0.585	0.586	0.586	0.586	0.585
	6	6	0.537	0.539	0.538	0.538	0.537
\log_{10} CPUtime			0.132	2.650	3.810	4.810	
Bias (x0.01)				0.277	0.210	0.144	0.008
RMSE (x0.01)				0.342	0.264	0.189	0.071
SD (x0.01)				0.011	0.010	0.014	

Table 2: Gittins index values found under the SBGIA for $n := n(i, 2) = 1, 3, 5$ and the Calibration method for the Bernoulli bandit. We set $K = 2$, $\gamma = 0.8$, $N = 35$, and set a tolerance of $\epsilon_\nu = 0.001$ in the stopping criterion (35) for determining the SBGIA. The computation (CPU) time denotes the time it took to calculate all values in the column and is measured in seconds. The bias, RMSE, and standard deviation (SD) are multiplied by 100, i.e., displayed in percentage points.

General	Ψ	$\kappa - \Psi$	Calibration	$n = 1$	$n = 3$	$n = 5$
	1	1	0.641	0.642	0.642	0.642
	1	2	0.443	0.446	0.445	0.444
	1	3	0.332	0.337	0.334	0.333
	1	4	0.263	0.268	0.266	0.265
	1	5	0.216	0.222	0.220	0.219
	1	6	0.183	0.189	0.187	0.186
	2	1	0.760	0.760	0.760	0.760
	2	2	0.590	0.591	0.591	0.590
	2	3	0.476	0.479	0.478	0.477
	2	4	0.398	0.401	0.399	0.399
	2	5	0.340	0.344	0.342	0.341
	2	6	0.296	0.300	0.299	0.298
	3	1	0.816	0.816	0.816	0.816
	3	2	0.671	0.673	0.672	0.672
	3	3	0.566	0.568	0.567	0.566
	3	4	0.487	0.489	0.489	0.488
	3	5	0.427	0.429	0.428	0.428
	3	6	0.379	0.382	0.380	0.380
	4	1	0.849	0.849	0.849	0.849
	4	2	0.725	0.725	0.725	0.725
	4	3	0.628	0.629	0.629	0.628
	4	4	0.552	0.554	0.553	0.553
	4	5	0.491	0.494	0.493	0.493
	4	6	0.443	0.445	0.444	0.444
	5	1	0.872	0.872	0.872	0.872
	5	2	0.762	0.763	0.763	0.762
	5	3	0.674	0.675	0.674	0.674
	5	4	0.602	0.603	0.603	0.602
	5	5	0.543	0.545	0.544	0.544
	5	6	0.494	0.496	0.496	0.495
	6	1	0.888	0.888	0.888	0.888
	6	2	0.790	0.791	0.791	0.790
	6	3	0.709	0.710	0.709	0.709
	6	4	0.641	0.642	0.642	0.642
	6	5	0.585	0.586	0.585	0.585
	6	6	0.537	0.538	0.538	0.537
\log_{10} CPUtime			0.132	3.810	3.950	4.170
Bias (x0.01)				0.210	0.113	0.059
RMSE (x0.01)				0.264	0.145	0.085
SD (x0.01)				0.010	0.016	0.013

Remark 4. *The length of the confidence interval for the Gittins index is in large part determined by the error bound for the optimal stopping value approximation given in Chen and Goldberg [2018]. The contribution of the optimal stopping approximation is equal to $2B(\delta, \xi)/c$ (Theorem 5), which can be seen as the bias of approximating $\nu_\sigma(\mathbf{s})$ by $\nu_M(\mathbf{s})$. Here c is the slope of $Z_t(\nu)$ in ν and $B(\delta, \xi)$ is a bound for the error induced by the approximation method introduced in Chen and Goldberg [2018]. The bound is similar to the radius of the confidence interval found under the Delta method when applied to sampled approximations to the truncated Gittins index $f_s^{-1}(0) = \nu_\sigma(\mathbf{s})$, which would be proportional to $1/f'_s(0)$. This implies that the bound could be sharp, given that the error bound $B(\delta, \xi)$ is sharp. Note that rescaling Z in (13) would not alter this radius, increasing the range of Z linearly increases the error according to Theorem 2, yielding the same confidence radius. The results in Section 5.2 indicate that the bound can be made tighter. For instance, for $\Psi = 1$, $\kappa = 2$, and $K = 2$, an absolute bias of $0.642 - 0.641 = 0.001$ is seen in Table 1. We have $R_u - R_\ell = 36/37 - 1/37 = 0.946$, $c = (1 - \gamma)/(2(R_u - R_\ell)(1 - \gamma^N)) = 0.106$, hence to get the theoretical bound for the bias less than 0.001 , we should at least have $B(\delta, \xi) = \xi + \delta \leq 0.106 \cdot 0.001/2 = 5.29 \cdot 10^{-5}$. Setting $\xi = \delta = 5.29 \cdot 10^{-5}/2$, following Chen and Goldberg [2018], we would need $K = \lfloor 1/\xi \rfloor = 378 \cdot 10^2$ and $n(i, k) = \lceil \log(2/\delta)/(2\xi^2) \rceil = 8.04 \cdot 10^9$ to obtain a bias of 0.001 for $\nu_\sigma(\mathbf{s})$. The main limiting factor in applying the theoretical bound in practice is hence the bound from Chen and Goldberg [2018] which could possibly be made more tight.*

5.2.3 Gaussian bandit

Let O_t^a be normally distributed with unknown mean θ_a , and known variance for each arm a . By scaling the mean and outcomes by the (known) standard deviation, we can equivalently assume $O_t^a \sim \mathcal{N}(\theta_a, 1)$, which implies that the family of outcome distributions is an exponential family with, letting ϕ denote the standard normal density,

$$\zeta \equiv \phi, \quad \eta(\theta_a) = \theta_a, \quad \psi(O) = O, \quad \rho(\theta_a) = \theta_a^2/2, \quad (46)$$

and μ the Lebesgue measure. We assume a $\mathcal{N}(\mu_0, v_a)$ prior on θ_a , hence $\kappa_t^a = 1/v_a + t$, $\Psi_t^a = \mu_0/v_a + \sum_{u=1}^t O_u^a$. In the following, we drop the superscript a from notation as we consider results for a single arm a only.

Observe from (46) that $O \mid (\Psi_t, \kappa_t) \sim \mathcal{N}(\Psi_t/\kappa_t, 1 + 1/\kappa_t)$, hence

$$R(\Psi_t, \kappa_t) = \mathbb{E}[O \mid \Psi_t, \kappa_t] = \Psi_t/\kappa_t,$$

and

$$R(\Psi_{t+1}, \kappa_{t+1}) \mid (R(\Psi_t, \kappa_t), \kappa_t) \sim \mathcal{N}(R(\Psi_t, \kappa_t), 1/(\kappa_t \kappa_{t+1})). \quad (47)$$

From Yao [2006] it holds that

$$\nu(\Psi_t, \kappa_t) = R(\Psi_t, \kappa_t) + \nu(0, \kappa_t).$$

It is hence sufficient to calculate the Gittins index for Gaussian rewards given that the initial sufficient statistic is zero. Observe from (47) that, starting from the initial state $(0, \kappa_t)$, the process $(R(\Psi_t, \kappa_{t+u}))_u$ is a Gaussian random walk starting at zero with normally distributed, zero-centered increments with variance $1/((\kappa_t + u - 1)(\kappa_t + u))$.

Let the stopping time σ be defined as in (15) for fixed $N \in \mathbb{N}$, and let $R_\ell = -L$, $R_u = L$ for fixed $L > 0$. We then have by Kolmogorov's inequality

$$\mathbb{P}_{(0, \kappa_t)}(\sigma < N) \leq \frac{\mathbb{E}[R(\Psi_{N-1}, \kappa_{t+N-1})^2]}{L^2} = \frac{1}{L^2} \sum_{u=1}^{N-1} 1/(\kappa_{t+u-1} \kappa_{t+u}) \leq \frac{1}{L^2 \kappa_t}.$$

We hence have $\mathbb{E}_{(0, \kappa_t)}[\gamma^\sigma] \leq \gamma^N + \frac{1}{L^2 \kappa_t}$. Note that

$$C(0, \kappa_t) = \sum_{u=0}^{\infty} \gamma^u \mathbb{E}_{(0, \kappa_t)} |R(\Psi_u, \kappa_{t+u})| = \sum_{u=1}^{\infty} \gamma^u \sqrt{\frac{2}{\pi} \left(\frac{1}{\kappa_t} - \frac{1}{\kappa_{t+u}} \right)} \leq \frac{\sqrt{2/\kappa_t}}{1 - \gamma}.$$

From Theorem 3, and as $\nu(\mathbf{S}_\sigma)^+ \leq (1 - \gamma)\mathbb{E}[C(0, \kappa_\sigma)] \leq \sqrt{2}$, it follows that

$$\nu(0, \kappa_t) - \nu_\sigma(0, \kappa_t) \leq \frac{\sqrt{2} \mathbb{E}_{(0, \kappa_t)}[\gamma^\sigma]}{1 - \mathbb{E}_{(0, \kappa_t)}[\gamma^\sigma]}, \quad (48)$$

The truncation error in (48) is smaller than $\epsilon_{\text{trunc}} > 0$ when, e.g.,

$$N \geq \log_\gamma \left(\epsilon_{\text{trunc}} / (2(\sqrt{2} + \epsilon_{\text{trunc}})) \right) \quad \text{and} \quad L \geq \sqrt{2(\sqrt{2} + \epsilon_{\text{trunc}}) / (\kappa_t \epsilon_{\text{trunc}})}. \quad (49)$$

Values of $R_u = -R_\ell = L$ and N such that the above inequalities are satisfied can then be used in (13).

Table 3 shows the Gittins index values found under the SBGIA and the Calibration method for the Gaussian bandit with $\gamma = 0.8$. For the SBGIA, we set $N = 39$, corresponding to a truncation error bound $\epsilon_{\text{trunc}} = 0.0005$ according to (48). For all states $(0, \kappa_t)$ the value of L was set to the lower bound in (49). We next set each $n(i, k) = 1$ and considered $K = 1, 2, 3$. The Gittins indices found under the Calibration method shown in Table 3 can also be derived from Gittins et al. [2011, Table 8.1]. As we only consider $\Psi = 0$, each state of the Gaussian bandit in Table 3 is denoted by κ . We show the log computation time at the bottom, as well as the bias, RMSE, and standard deviation in percentage points. The rightmost column of the table shows an estimate of the limit as $K \rightarrow \infty$, found by performing ordinary least squares on the columns $K = 1, 2, 3$ with the line $\nu(s) = a + b/(K + 1)$.

Table 3 shows that, as in Table 1, the Gittins index is overestimated for $K = 1$, and the values for the SBGIA, as well as the error measures, decrease in K . The estimates for the Gaussian bandit show larger errors than those for the Bernoulli bandit. Possibly due to the continuity in the support of the rewards, which induces a larger variance in the sampled paths. The computation times for the Gaussian bandit are also approximately ten times larger than those for the Bernoulli bandit. The computation time shown in Table 3 is similar for $K = 1, 2$, and increases tenfold when going from $K = 2$ to $K = 3$. The computation time of the Calibration method is comparable to the computation time for $K = 1$, indicating that the Gaussian bandit with known variance is already a hard problem to solve under the Calibration method. The low average standard deviation in Table 3 indicates that the estimates are consistent over different runs. The estimates of the limits in the rightmost column again show a better quality than those for finite K , often giving the value of the Gittins index with an error of 0.001 for $\kappa \geq 4$.

Table 4 shows values of the SBGIA obtained when setting $K = 2$ and varying $n(i, 2)$ for the Gaussian bandit with unit variance. As in Table 2, it is seen that the errors decrease faster in $n(i, k)$ for K fixed than vice versa, in agreement with the proposal in Remark 3.

Table 3: Gittins index values found under the SBGIA for $K = 1, 2, 3$, the Calibration method, and an estimate for the limit as $K \rightarrow \infty$ for the Gaussian bandit (unit variance). For columns $K = 1, \dots, 3$ we set each $n(i, k) = 1$ $\gamma = 0.8$, $N = 39$, and set a tolerance of $\epsilon_\nu = 0.001$ in the stopping criterion (35) for determining the SBGIA. The computation (CPU) time denotes the time it took to calculate all values in the column and is measured in seconds. The bias, RMSE, and standard deviation (SD) are multiplied by 100, i.e., displayed in percentage points.

General	κ	Calibration	$K = 1$	$K = 2$	$K = 3$	Est. limit
	1	0.505	0.526	0.520	0.520	0.513
	2	0.308	0.329	0.323	0.320	0.312
	3	0.226	0.245	0.239	0.237	0.229
	4	0.179	0.196	0.191	0.188	0.180
	5	0.149	0.164	0.160	0.157	0.150
	6	0.128	0.142	0.138	0.135	0.128
	7	0.112	0.125	0.121	0.119	0.113
	8	0.100	0.112	0.108	0.106	0.101
	9	0.090	0.101	0.098	0.096	0.091
	10	0.082	0.092	0.089	0.087	0.083
	20	0.043	0.050	0.048	0.047	0.044
	30	0.029	0.034	0.033	0.032	0.030
	40	0.022	0.026	0.025	0.025	0.023
	50	0.018	0.021	0.020	0.020	0.018
\log_{10} CPUtime		4.050	3.860	4.320	5.730	
Bias (x0.01)			1.250	0.871	0.725	0.172
RMSE (x0.01)			1.380	0.970	0.818	0.263
SD (x0.01)			0.030	0.031	0.059	

Table 4: Gittins index values found under the SBGIA for $n := n(i, 2) = 1, 3, 5$ and the Calibration method for the Gaussian bandit (unit variance). We set $K = 2$, $\gamma = 0.8$, $N = 39$, and set a tolerance of $\epsilon_\nu = 0.001$ in the stopping criterion (35) for determining the SBGIA. The computation (CPU) time denotes the time it took to calculate all values in the column and is measured in seconds. The bias, RMSE, and standard deviation (SD) are multiplied by 100, i.e., displayed in percentage points.

General	κ	Calibration	$n = 1$	$n = 3$	$n = 5$
	1	0.505	0.520	0.514	0.512
	2	0.308	0.323	0.317	0.316
	3	0.226	0.239	0.235	0.232
	4	0.179	0.191	0.188	0.185
	5	0.149	0.160	0.157	0.155
	6	0.128	0.138	0.135	0.133
	7	0.112	0.121	0.119	0.116
	8	0.100	0.108	0.107	0.104
	9	0.090	0.098	0.096	0.094
	10	0.082	0.089	0.088	0.085
	20	0.043	0.048	0.047	0.046
	30	0.029	0.033	0.032	0.031
	40	0.022	0.025	0.025	0.024
	50	0.018	0.020	0.020	0.019
\log_{10} CPUtime		4.050	4.320	4.710	4.880
Bias (x0.01)			0.871	0.644	0.437
RMSE (x0.01)			0.970	0.698	0.481
SD (x0.01)			0.031	0.015	0.010

5.3 Gaussian random effects bandit

This section compares the performance of a policy based on the SBGIA to that of policies Thompson sampling and Bayes-UCB in case each arm describes the posterior under a Gaussian random effects model. This multi-armed bandit model was not found in literature.

In the Gaussian random effects bandit model, it is assumed that there an additional factor that induces heterogeneity within each of the A distributions of choice. The factor induces multiple clusters to which the outcomes are assigned. Outcomes assigned to the same cluster have the same expected value, which deviates from the overall expected value for the distribution. As each deviation is induced by a common factor, the deviations are sampled from a common distribution. The assumed model is (hence) an independent mixed effects model (intercept and random effects) for the outcomes under each of the A distributions.

For $d \in \mathbb{N}$, let $\mathbf{C}_{a,t} \in \{0, 1\}^d$ be a vector denoting cluster assignment for outcome O_t^a such that $\sum_{i=1}^d C_{a,t,i} = 1$. We assume for all t that

$$O_t^a = \theta_a + \mathbf{C}_{a,t}^\top \mathbf{u}_a + \epsilon_{a,t}, \quad (50)$$

where, independently,

$$\epsilon_{a,t} \sim \mathcal{N}(0, v_a^{(1)}) \text{ and } u_{a,i} \sim \mathcal{N}(0, v_a^{(2)}).$$

The set of model parameters for each arm a hence consists of $(\theta_a, \mathbf{u}_a, v_a^{(1)}, v_a^{(2)})$, and no parameters are shared between the arms. The process of cluster assignment $(\mathbf{C}_{a,t})_t$ is assumed predictable, i.e., all cluster assignments are known prior to assignment to the arm. The prior specification is as follows, we assume a normal $\mathcal{N}(\theta_0, \sigma_0^2)$ prior on each θ_a , an inverse-gamma $\mathcal{IG}(\alpha_0, \beta_0)$ prior on each $v_a^{(1)}$, and an $\mathcal{IG}(\alpha_1, \beta_1)$ prior on each $v_a^{(2)}$.

The above data model and prior specification lead to an analytic expression for the full conditional distribution of each parameter, and hence an efficient Gibbs sampling procedure such as the one in Wang et al. [1993] can be constructed. This Gibbs sampling procedure can then be included in a sequential Markov chain Monte Carlo method (Chopin [2002]) in order to efficiently update approximations of the posterior distribution Π_t^a upon sampling a new observation O_{t+1}^a . For the sequential Markov chain Monte Carlo method the set of observations $(O_t^a)_t$ leads to a collection of samples $(\theta_{a,i,t}, \mathbf{u}_{a,i,t}, v_{a,i,t}^{(1)}, v_{a,i,t}^{(2)})_{i=1}^{d_2}$ of particles and weights $(w_{a,i,t})_{i=1}^{d_2}$ such that $\sum_{i=1}^{d_2} w_{a,i,t} = 1$ and, denoting with δ_x the Dirac measure at x ,

$$\hat{\Pi}_t^a = \sum_{i=1}^{d_2} w_{a,i,t} \delta_{(\theta_{a,i,t}, \mathbf{u}_{a,i,t}, v_{a,i,t}^{(1)}, v_{a,i,t}^{(2)})} \approx \Pi_t^a. \quad (51)$$

Based on this approximation to the posterior, we consider three policies for the Bayesian multi-armed bandit:

- **SBGIAP:** Determine the SBGIA by sampling future approximations $\hat{\Pi}_t^a$ to the posterior Π_t^a from the Markov chain that approximates (41) with transition kernel

$$\mathbb{P}_{\hat{\Pi}_t^a}(E) = \int_{\mathcal{O}} \mathbb{Q}(\hat{\Pi}_{t+1}^a(\cdot | O, \hat{\Pi}_t^a) \in E) p(O | \hat{\Pi}_t^a) \mu(dO) \quad \forall E \in \mathcal{M},$$

where \mathbb{Q} denotes the measure on approximate posteriors induced by a sequential Monte Carlo step using d_2 particles, given the current approximation to the posterior $\hat{\Pi}_t^a$ and the sampled outcome O . The reward function for the Markov chain is given by the posterior mean under the empirical distribution

$$R(\hat{\Pi}_t^a) = \sum_{i=1}^{d_2} w_{a,i,t} (\theta_{a,i,t} + \mathbf{C}_{a,t}^\top \mathbf{u}_{a,i,t}).$$

As in Theorem 6, the SBGIAP now chooses $A_t = \arg \max_{a \in [A]} \nu_M(\hat{\Pi}_t^a)$, where ν_m is determined as the M -th

iterate of (18) for a choice of K, N, n, M . To decrease the numerical burden, the approximated posteriors $\hat{\Pi}_{t+1}^a, \dots, \hat{\Pi}_{t+N}^a$ in the SBGIAP are based on $d_3 \leq d_2$ samples after sampling the first observation O , by sampling d_3 particles to continue with from the initial distribution $\hat{\Pi}_t^a$.

- **Thompson sampling (Thompson [1933]):** Sample $i \sim \text{Categorical}_{d_2}(\mathbf{w}_{a,t})$ and set $\eta_{a,t} = \theta_{a,i,t} + \mathbf{C}_{a,t+1}^\top \mathbf{u}_{a,i,t}$. Choose $A_t = \arg \max_{a \in [A]} \eta_{a,t}$.
- **Bayesian upper confidence bound (Bayes-UCB) (Kaufmann et al. [2012a]):** Set

$$\eta_{a,t} = \hat{q}((\theta_{a,i,t} + \mathbf{C}_{a,t+1}^\top \mathbf{u}_{a,i,t})_{i=1}^{d_2}, 1 - (t \log(T)^6)^{-1}),$$

where $\hat{q}((\rho_i, w_i)_i, 1 - \alpha)$ is the empirical $1 - \alpha$ quantile given the samples ρ_i and weights w_i , and where T is the total sample size of the experiment. Choose $A_t = \arg \max_{a \in [A]} \eta_{a,t}$.

The total discounted rewards found under the policies are compared using a simulation study. We note that the policies Bayes-UCB and Thompson sampling, unlike the SBGIAP, are not tuned to a specific discount factor γ . Other Bayesian bandit policies tuned to a specific discount factor are not known from literature, and introducing them in the current paper would deviate attention from the SBGIAP. Policies Bayes-UCB and Thompson sampling have good performance guarantees for undiscounted reward (Kaufmann et al. [2012b,a]), hence in order to have a fair comparison, we compare the performance of the three policies when higher discount factors $\tilde{\gamma} \in \{0.8, 0.9, 0.99\}$ are used to determine the total discounted reward, while we tune the SBGIAP to $\gamma = 0.8$. An outperformance over Bayes-UCB and Thompson sampling in terms of Bayesian total discounted reward (with discount factor $\tilde{\gamma} = \gamma$) is expected for the Gittins index policy, as it exactly maximizes this quantity. The SBGIA is however an approximation to the Gittins index, and hence outperformance for the SBGIAP in terms of the Bayesian total discounted reward is not guaranteed, furthermore there is no guarantee that a policy based on the Gittins index tuned to $\gamma = 0.8$ also outperforms other policies for other discount factors $\tilde{\gamma}$. Hence, it is interesting to compare the performance of the policy using a simulation study.

For the simulation study, we set $\theta_0 = 0$, $\sigma_0^2 = 1$, $\alpha_0 = 13$, $\beta_0 = 12$, $\alpha_1 = 6$, $\beta_1 = 10$. In order to approximate the Bayesian total discounted reward, the parameters were sampled from the resulting prior distributions for each simulation. The sample size of the simulation study was set to $T = 300$, the number of clusters d was set to 3, and the number of arms A was set to 3. The cluster assignments \mathbf{C}_a were sampled uniformly for each arm a . Given the sampled parameters and cluster assignments, the vector of observations $\mathbf{O}^a \in \mathbb{R}^{300}$ were sampled according to model (50). Given the data and cluster assignments, a sequence of weights $(\mathbf{w}_{a,t})_{t=1}^{300}$ ($d_2 = 100$) and particles $(\theta_{a,t}, \sigma_{a,t}^2, \tau_{a,t}, \mathbf{u}_{a,t})_{t=1}^{300}$ was generated for each arm using sequential Markov chain Monte Carlo sampling (starting with a sample from the prior) using 5 Gibbs sampler iterations in each Markov chain Monte Carlo step. Each algorithm then determined an interleaving of these independent Markov chain samples, where for the SBGIAP, we set $K = 1$, $N = 25$, each $n(i, k) = 1$, $M = 100$, and $d_3 = 3$. The above procedure, resulting in an interleaving of the sampled Markov chain $(\mathbf{w}_{a,t}, \theta_{a,t}, \mathbf{u}_{a,t}, \sigma_{a,t}^2, \tau_{a,t})_{t=1}^{300}$ for each algorithm described above, is then repeated independently 2500 times to approximate the Bayesian total discounted reward. This procedure took about 10 days on a computer with 32 cores.

Figure 1 shows results for the SBGIAP with $\gamma = 0.8$. Differences (averaged over 2500 simulations) between Bayesian total discounted reward for the SBGIAP vs. Bayes-UCB (solid, blue) and Thompson sampling vs. Bayes-UCB (dotted, red) are shown for discount factors $\tilde{\gamma} = 0.8, 0.9$, and 0.99 , along with a point-wise 95% bootstrapped confidence

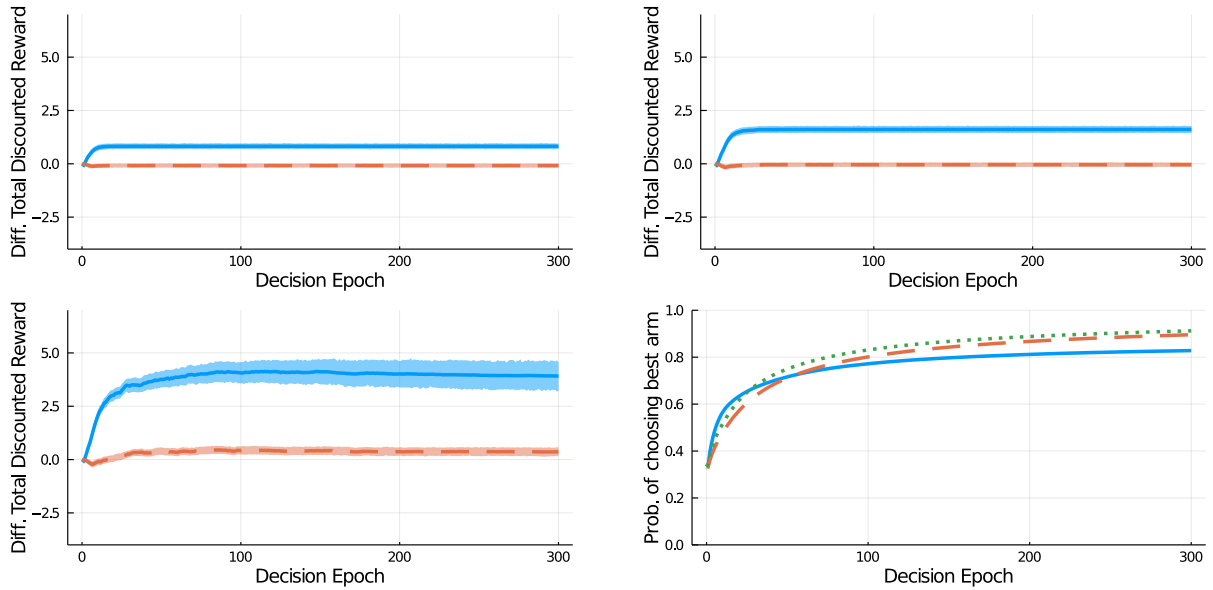


Figure 1: Results for the SBGIAP for the Gaussian random effects Bayesian multi-armed bandit model. The difference in running total discounted reward for a discount factor of 0.8 (top-left), 0.9 (top-right), and 0.99 (bottom-left) is shown for the SBGIAP vs. Bayes-UCB (solid) and Thompson sampling vs. Bayes-UCB (dashed). The running frequency of choosing the best arm for the SBGIAP (solid), Thompson sampling (dashed) and Bayes-UCB (dotted) is shown on the bottom-right.

interval for the mean. The total discounted rewards at each decision epoch are calculated as the discounted sum of the expected rewards ($\mathbb{E}[O_t] = \theta_{A_t} + C_{A_t,t} \mathbf{u}_{A_t}$) from the initial decision epoch up to that time point.

Figure 1 shows that the SBGIAP significantly outperforms Bayes-UCB and Thompson sampling with a final difference in average total discounted reward of about 1, 2 and 4 for discount factors 0.8, 0.9, and 0.99 respectively.

The average (undiscounted) frequency of choosing the best arm is calculated as the frequency each algorithm chose $a_t^* = \arg \max_a \theta_a + C_{a,t}^\top \mathbf{u}_a$ at each decision epoch. It is seen that while the SBGIAP reaches a frequency of 60% of optimal pulls early on, the other strategies end up at a higher frequency of optimal pulls. This might be a result of the low discount factor used for constructing the policy, indicating that short-term gains are preferred over long-term ones. The bottom-right graph is in line with the other graphs in Figure 1 as, when considering the sum of discounted rewards starting at the initial state, making good choices early on has a larger benefit than outperformance in the long run.

We conclude that with the SBGIAP, significant outperformance in terms of total discounted reward with respect to state-of-the-art policies can be attained for more complex models than usually considered in bandit literature.

6 Discussion and conclusion

In this paper, we have proposed the sampling-based Gittins index approximation (SBGIA). In Section 4, the SBGIA was introduced, and a general error bound was shown for the Gittins index obtained when truncating the horizon and support for the rewards using a stopping time, which can be viewed as an extension of the results presented in Wang [1997]. Next, finite and asymptotic convergence results were shown for the SBGIA. Using the finite-time convergence result it is possible to obtain a confidence interval for the Gittins index which holds under a finite number of stochastic approximation samples. Next, it was shown in Theorem 6 that by making explicit choices for the width and safety level δ_2 of this confidence interval, the policy choosing the largest Gittins index estimate is an ϵ -optimal policy for the Bayesian multi-armed bandit problem. For both the Bernoulli and Gaussian bandits, the SBGIA was seen to yield a good approximation to the Gittins index, increasing in quality with the number of nested simulations and the truncation parameter K , and showing the best approximation when estimating the limit as K goes to infinity. The results indicated that an efficient strategy for Gittins index approximation using SBGIA is to first increase the number of nested simulations, and then increase the truncation parameter K , until no significant differences in the estimate are

seen for both steps. The SBGIA can be applied even in cases where the actual transition kernel is unknown, but where samples from an approximation to the transition kernel can be generated. An example of this was seen in Section 5.3, where samples from the approximate posterior were generated using sequential Monte Carlo sampling. In this case, the SBGIAP was seen to outperform the state-of-the-art policies Bayes-UCB and Thompson sampling in terms of Bayesian total discounted reward.

The SBGIA can be applied to any family of alternative bandit processes. For example, to compute the Gittins index approximations in Section 5.2, only three things must be altered for each bandit, namely the transition kernel, the reward function, and calculation of the bounds R_ℓ, R_u . In contrast, for the Calibration method, an additional requirement is that the reward support also has to be discretized and the change in state space leads to a reformulation of the backward induction step. The benefit of a method that works in general, is that there is more flexibility in the model choice when basing treatment allocation on the Gittins index, as there is no increased difficulty in implementing the calculation method when assuming a more elaborate model for the data. Another benefit is that less expert knowledge is necessary for Gittins index approximation. It might be an interesting idea to have a software library where practitioners only have to input functions that calculate, e.g., the posterior mean, after which the package calculates the SBGIA. It is furthermore useful to have a method that does not assume known transition probabilities, as in many real-life cases the posterior distribution cannot be calculated in closed form because of the high dimensionality of the model or when the assumed prior is nonconjugate.

In Section 5.3, we evaluated the performance of the SBGIA policy (SBGIAP) for a novel random effects bandit problem. The SBGIAP was defined based on an approximation of the Markov chain $(\Pi_t^a)_t$ describing the evolution of the posterior distribution, based on sequential Markov chain Monte Carlo. In future research, it would be interesting to investigate how finite-time convergence results for Markov chains (e.g., Rosenthal [1995]) can be used to construct finite-time error bounds for the SBGIA in these situations. In this paper, the SBGIA was evaluated for a number of Bayesian multi-armed bandit problems. In the case of Gaussian outcomes, the Gittins index was shown to result in near-optimal frequentist undiscounted regret (Lattimore [2016]). If this result is shown to hold in general, the confidence interval presented in Theorem 5 ensures that we have a method that can approximate, up to arbitrary precision, a near-optimal policy, in terms of undiscounted frequentist regret, for the multi-armed bandit problem.

References

- S Aalto, U Ayesta, and R Righter. Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25(3):269–288, 2011.
- Krishna B Athreya and Soumendra N Lahiri. *Measure theory and probability theory*. Springer; New York, 2006.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, 2002.
- S Boodaghians, F Fusco, P Lazos, and S Leonardi. Pandora’s Box Problem with Order Constraints. *Mathematics of Operations Research*, 48(1):1–602, 2023.
- Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press; Cambridge, 2008.
- Bruce M Brown. Martingale central limit theorems. *The Annals of Mathematical Statistics*, 42(1):59–66, 1971.
- Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
- Jhelum Chakravorty and Aditya Mahajan. Multi-armed bandits, Gittins index, and its calculation. *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods*, 2:416–435, 2014.
- Y. Chen and D. Goldberg. Beating the curse of dimensionality in options pricing and optimal stopping. *arXiv preprint, arXiv:1807.02227*, 2018. Available at <https://arxiv.org/abs/1807.02227v2>, (accessed 8-6-2023).
- Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.
- Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.
- Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.
- James Edwards. Practical calculation of Gittins indices for multi-armed bandits. *arXiv preprint arXiv:1909.05075*, 2019. Available at <https://arxiv.org/abs/1909.05075> (accessed 12-6-2023).
- Evan Fisher. An upper class law of the iterated logarithm for supermartingales. *Sankhyā: The Indian Journal of Statistics, Series A*, 48:267–272, 1986.

- Subhashis Ghosal and Aad Van der Vaart. Fundamentals of nonparametric Bayesian inference. Cambridge University Press; Cambridge, 2017.
- Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press; Routledge, 1995.
- John Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.
- John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons: Hoboken, 2011.
- Kevin D Glazebrook. On the evaluation of suboptimal strategies for families of alternative bandit processes. *Journal of Applied Probability*, 19(3):716–722, 1982.
- Kevin D Glazebrook. Optimal strategies for families of alternative bandit processes. *IEEE Transactions on Automatic Control*, 28(8):858–861, 1983.
- Kevin D Glazebrook and R Minty. A generalized Gittins index for a class of multiarmed bandits with general resource requirements. *Mathematics of Operations Research*, 34(1):26–44, 2009.
- Emilie Kaufmann. On Bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2): 842–865, 2018.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, 2012a.
- Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 199–213. Springer; Berlin, Heidelberg, 2012b.
- Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014. available at <https://arxiv.org/abs/1402.6028> (accessed 8-6-2023).
- T L_ Lai and Herbert Robbins. Limit theorems for weighted sums and stochastic approximation processes. *Proceedings of the National Academy of Sciences*, 75(3):1068–1070, 1978.
- T L_ Lai and Herbert Robbins. Adaptive design and stochastic approximation. *The Annals of Statistics*, 7(6):1196–1221, 1979.
- Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3): 1091–1114, 1987.
- Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.
- T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press; Cambridge, 2020.
- Tor Lattimore. Regret Analysis of the Finite-Horizon Gittins Index Strategy for Multi-Armed Bandits. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1214–1245, 23–26 Jun 2016.
- Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In A.O. Hero, D. Castañón, D. Cochran, and K. Kastella, editors, *Foundations and Applications of Sensor Management*, Signals and Communication Technology, chapter 6, pages 121–151. Springer: New York, 2008.
- B Neuenschwander, S Weber, H Schmidli, and A O’Hagan. Predictively consistent prior effective sample sizes. *Biometrics*, 76(2):578–587, 2020.
- Martin L Puterman. Markov decision processes. In D.P. Heyman and M.J. Sobel, editors, *Handbooks in Operations Research and Management Science*, volume 2, chapter 8, pages 331–434. Elsevier; Amsterdam, 1990.
- Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.
- David S Robertson, Kim May Lee, Boryana C López-Kolkovska, and Sofía S Villar. Response-adaptive randomization in clinical trials: from myths to practical considerations. *Statistical science*, 38(2):185–208, 2023.
- J S Rosenthal. Convergence rates for markov chains. *Siam Review*, 37(3):387–405, 1995.

- Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning*, 12(1-2): 1–286, 2019.
- William 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.
- S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science*, 30(2):199–215, 2015a.
- Sofía S. Villar, James Wason, and Jack Bowden. Response-adaptive randomization for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics*, 71(4):969–978, 2015b.
- CS Wang, JJ Rutledge, and D Gianola. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution*, 25:41–62, 1993.
- You-Gan Wang. Error bounds for calculation of the Gittins indices. *Australian Journal of Statistics*, 39(2):225–233, 1997.
- Richard Weber. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.
- Richard Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3): 637–648, 1990.
- Yi-Ching Yao. Some results on the Gittins index for a normal reward process. In *Time Series and Related Topics*, volume 52 of *Lecture Notes–Monograph Series*, pages 284–294. Institute of Mathematical Statistics, 2006.

A Proofs of theorems

Theorem 3. *Under Assumption 1 there is a unique real number $\nu_\sigma(\mathbf{s})$ attaining the supremum*

$$\sup \left\{ \nu : \inf_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_{\mathbf{s}}[\tilde{Z}_\tau(\nu)] \leq 0 \right\}.$$

In fact, $\nu_\sigma(\mathbf{s})$ is the unique root of

$$f_{\mathbf{s}}^\sigma : \nu \mapsto \inf_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_{\mathbf{s}}[\tilde{Z}_\tau(\nu)].$$

The infimum in the condition is attained for some $\tau_\sigma(\nu, \mathbf{s}) \in \mathcal{T}_\sigma$. Furthermore, with $x^+ = \max(x, 0)$ for $x \in \mathbb{R}$, it holds that

$$0 \leq \nu(\mathbf{s}) - \nu_\sigma(\mathbf{s}) \leq \frac{\mathbb{E}_{\mathbf{s}}[\gamma^\sigma \nu(\mathbf{S}_\sigma)^+]}{1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma]}.$$

Proof. Let $(\tilde{\mathbf{S}}_t)_t$ be a Markov chain with states $\tilde{\mathbf{S}}_t = (\mathbf{S}_u)_{u=0}^t$. Let $\tilde{R}(\tilde{\mathbf{S}}_t) = R(\tilde{\mathbf{S}}_t)\mathbb{I}(t < \sigma)$. The pair $(\tilde{\mathbf{S}}, \tilde{R})$ defines a Markov reward process as $\mathbb{I}(t < \sigma)$ is a function of $\tilde{\mathbf{S}}_t$. As the filtrations and hence the set of stopping times generated by \mathbf{S} and $\tilde{\mathbf{S}}$ are the same, we have

$$\sup_{\tau \in \mathcal{T}} \mathbb{E}_{\mathbf{s}} \left[\sum_{t=0}^{\tau-1} \gamma^t (\tilde{R}(\tilde{\mathbf{S}}_t) - \nu) \right] = \sup_{\tau \in \mathcal{T}} \mathbb{E}_{\mathbf{s}} \left[\sum_{t=0}^{\tau \wedge \sigma - 1} \gamma^t (R(\mathbf{S}_t) - \nu) \right] = \sup_{\tau \in \mathcal{T}_\sigma} \mathbb{E}_{\mathbf{s}} \left[\sum_{t=0}^{\tau-1} \gamma^t (R(\mathbf{S}_t) - \nu) \right]. \quad (52)$$

Note that as Assumption 1 holds for the Markov reward process defined by (\mathbf{S}, R) , it also holds for the Markov reward process defined by $(\tilde{\mathbf{S}}, \tilde{R})$. Hence from Lattimore and Szepesvári [2020] there is a unique value for ν , denoted by $\nu_\sigma(\mathbf{s})$, such that the left-hand side, hence the right-hand side of (52) is zero. As $f_{\mathbf{s}}^\sigma$ is a scaling of (52), the first two statements of the theorem are proven. From Lattimore and Szepesvári [2020], we also have that the supremum in (52) is attained for an unique stopping time $\tau(\mathbf{s}, \nu)$, proving the third statement.

Now, we bound the approximation error from above (the lower bound holds trivially):

$$\begin{aligned}
\nu(\mathbf{s}) - \nu_\sigma(\mathbf{s}) &= \sup_{\tau \in \mathcal{T}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t]} - \sup_{\tau \in \mathcal{T}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau \wedge \sigma-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t]} \\
&\leq \sup_{\tau \in \mathcal{T}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t R(\mathbf{S}_t)] - \mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau \wedge \sigma-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t]} \\
&= \sup_{\tau \in \mathcal{T}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=\sigma \wedge \tau}^{\tau-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t]} = \sup_{\substack{\tau \in \mathcal{T} \\ \tau \geq \sigma}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=\sigma}^{\tau-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\tau-1} \gamma^t]} \\
&\leq \sup_{\substack{\tau \in \mathcal{T} \\ \tau \geq \sigma}} \frac{\mathbb{E}_{\mathbf{s}}[\sum_{t=\sigma}^{\tau-1} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{\sigma-1} \gamma^t]} = \sup_{\substack{\tau \in \mathcal{T} \\ \tau \geq \sigma}} \frac{(1-\gamma)\mathbb{E}_{\mathbf{s}}[\gamma^\sigma \mathbb{E}_{\mathbf{s}}[\sum_{t=\sigma}^{\tau-1} \gamma^{t-\sigma} R(\mathbf{S}_t) | \mathcal{F}_\sigma]]}{1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma]} \\
&\leq \sup_{\substack{\tau \in \mathcal{T} \\ \tau \geq \sigma}} \frac{\mathbb{E}_{\mathbf{s}} \left[\gamma^\sigma \left(\frac{\mathbb{E}_{\mathbf{s}_\sigma}[\sum_{t=0}^{\tau-\sigma} \gamma^t R(\mathbf{S}_t)]}{\mathbb{E}_{\mathbf{s}_\sigma}[\sum_{t=0}^{\tau-\sigma} \gamma^t]} \right)^+ \right]}{1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma]} \leq \frac{\mathbb{E}_{\mathbf{s}}[\gamma^\sigma \nu(\mathbf{S}_\sigma)^+]}{1 - \mathbb{E}_{\mathbf{s}}[\gamma^\sigma]}.
\end{aligned}$$

The second to last inequality above holds by the strong Markov property and as $(1-\gamma) \leq (1-\gamma)/(1-\mathbb{E}_{\mathbf{s}_\sigma}[\gamma^{\tau-\sigma+1}])$ almost surely. The ‘‘positive part’’ of the Gittins index in the last term above comes from the fact that we can take $\tau = \sigma$ in the seventh term. \square

Theorem 4. For $\omega^* \in \mathbb{R}$, initial point $\omega_0 \in \mathbb{R}$, and martingale difference sequence (21), let the sequence $(\omega_m)_m$ be defined as

$$\omega_{m+1} = \omega_m - \alpha_m(f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*) + \epsilon_m).$$

Assume the step-size sequence is such that $\sup_m \alpha_m \leq 2(R_u - R_\ell)$ almost surely. Then for all m

$$\mathbb{E}[(\omega^* - \omega_{m+1})^2] \leq \mathbb{E}[(1 - c\alpha_m)^2 (\omega^* - \omega_m)^2] + \mathbb{E}[\alpha_m^2] \mathbb{E}[\epsilon_m^2].$$

Proof. Let $\mathcal{F}_m^c = \sigma(\epsilon_0, \dots, \epsilon_m)$ such that ω_m and α_m are \mathcal{F}_{m-1}^c -measurable. Then, as $\mathbb{E}[\epsilon_m | \mathcal{F}_{m-1}^c] = 0$ for $m \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[(\omega^* - \omega_{m+1})^2 | \mathcal{F}_{m-1}^c] &= \mathbb{E}[(\omega^* - \omega_m + \alpha_m(f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*)) + \alpha_m \epsilon_m)^2 | \mathcal{F}_{m-1}^c] \\
&= (\omega^* - \omega_m + \alpha_m(f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*)))^2 + \alpha_m^2 \mathbb{E}[\epsilon_m^2].
\end{aligned} \tag{53}$$

Now, by (25), letting $\eta(\omega_1, \omega_2) = 1 + \left(\frac{1-\gamma}{1-\gamma^N} - 1\right) \mathbb{1}_{[\omega_2 \geq \omega_1]}$ we have

$$\left(1 - \frac{\alpha_m \eta(\omega^*, \omega_m)}{2(R_u - R_\ell)}\right) (\omega^* - \omega_m) \leq \omega^* - \omega_m + \alpha_m(f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*)) \leq \left(1 - \frac{\alpha_m \eta(\omega_m, \omega^*)}{2(R_u - R_\ell)}\right) (\omega^* - \omega_m)$$

Hence

$$\begin{aligned}
&|\omega^* - \omega_m + \alpha_m(f_{\mathbf{s}}(\omega_m) - f_{\mathbf{s}}(\omega^*))| \\
&\leq \max\left(\left(1 - \frac{\alpha_m \eta(\omega^*, \omega_m)}{2(R_u - R_\ell)}\right) (\omega_m - \omega^*), \left(1 - \frac{\alpha_m \eta(\omega_m, \omega^*)}{2(R_u - R_\ell)}\right) (\omega^* - \omega_m)\right) \\
&= \left(1 - \frac{\alpha_m(1-\gamma)}{2(R_u - R_\ell)(1-\gamma^N)}\right) |\omega^* - \omega_m|.
\end{aligned}$$

The last line above follows as $\eta \in [0, 1]$, hence $\left(1 - \frac{\alpha_m \eta(x, y)}{2(R_u - R_\ell)}\right) \geq 0$ for all x, y by the assumptions on $(\alpha_m)_m$ and the maximum will be attained at the first argument if $\omega_m \geq \omega^*$ where $\eta(\omega^*, \omega_m) = \frac{1-\gamma}{1-\gamma^N}$ and the maximum will be attained at the second argument if $\omega^* \geq \omega_m$ where $\eta(\omega_m, \omega^*) = \frac{1-\gamma}{1-\gamma^N}$. Then, continuing (53), we have:

$$\mathbb{E}[(\omega^* - \omega_{m+1})^2 | \mathcal{F}_{m-1}^c] \leq \left(1 - \frac{\alpha_m(1-\gamma)}{2(R_u - R_\ell)(1-\gamma^N)}\right)^2 (\omega^* - \omega_m)^2 + \alpha_m^2 \mathbb{E}[\epsilon_m^2].$$

To conclude, taking expectations, we have

$$\mathbb{E}[(\omega^* - \omega_{m+1})^2] \leq \mathbb{E}[(1 - c\alpha_m)^2 (\omega^* - \omega_m)^2] + \mathbb{E}[\alpha_m^2] \mathbb{E}[\epsilon_m^2].$$

where $c = \frac{1-\gamma}{2(R_u - R_\ell)(1-\gamma^N)}$. \square

Lemma 1. *Let $\sup_m \alpha_m \leq M_\alpha < \infty$ almost surely. We have almost surely that for all m*

$$\nu_m(\mathbf{s}) \in [R_\ell - M_\alpha/2, R_u + M_\alpha/2].$$

Proof. Note that if $\nu > R_u$ it holds that irrespective of the path \mathbf{S}

$$\min_{u \in [N]} Z_{i,u,n}^{(1)}(\nu) = \min_{u \in [N]} c \sum_{w=0}^{(u \wedge \sigma_H) - 1} \gamma^w (\nu - R(\mathbf{S}_w)) = c(\nu - R(\mathbf{S}_0)) = Z_{i,1,n}^{(1)}(\nu) > 0.$$

For $k = 2$, we hence have irrespective of the path \mathbf{S} , that when $\nu > R_u$

$$\begin{aligned} \min_{u \in [N]} Z_{i,u,n}^{(2)}(\nu) &= \min_{u \in [N]} Z_{[i,1],u,n}^{(1)}(\nu) - \frac{1}{n(\mathbf{i}, 2)} \sum_{j=2}^{n(\mathbf{i}, 2) + 1} \left(\min_{w \in [N]} Z_{[i,j],w,n}^{(1)}(\nu) \Big| \{\mathbf{S}_{[0,u]}^{[i,j]} = \mathbf{S}_{[0,u]}^i\} \right) \\ &= \min_{u \in [N]} Z_{[i,1],u,n}^{(1)}(\nu) - \left(Z_{[i,j],1,n}^{(1)}(\nu) \Big| \{\mathbf{S}_{[0,u]}^{[i,j]} = \mathbf{S}_{[0,u]}^i\} \right) = 0. \end{aligned}$$

From this, we see that for $k > 2$

$$\min_{u \in [N]} Z_{i,u,n}^{(k)}(\nu) = \min_{u \in [N]} Z_{[i,1],u,n}^{(k-1)}(\nu) - \frac{1}{n(\mathbf{i}, k)} \sum_{j=2}^{n(\mathbf{i}, k) + 1} \left(\min_{w \in [N]} Z_{[i,j],w,n}^{(k-1)}(\nu) \Big| \{\mathbf{S}_{[0,u]}^{[i,j]} = \mathbf{S}_{[0,u]}^i\} \right) = 0 - 0 = 0.$$

Hence for $\nu > R_u$ we have almost surely

$$\begin{aligned} V_{\mathbf{s}}^{(K)}(\nu) &= \max \left(-1/2, \min \left(1/2, \frac{1}{n(K, K)} \sum_{j=1}^{n(K, K)} \sum_{k=1}^K \min_{u \in [N]} Z_{[k,j],u,n}^{(k)} \right) \right) \\ &= \max \left(-1/2, \min \left(1/2, \frac{1}{n(K, K)} \sum_{j=1}^{n(K, K)} Z_{[1,j],1,n}^{(1)} \right) \right) > 0. \end{aligned}$$

Similarly, observe that $V_{\mathbf{s}}^{(K)}(\nu) < 0$ almost surely if $\nu < R_\ell$. Hence by (18) we have almost surely for all m that

$$\nu_m \in [R_\ell - M_\alpha/2, R_u + M_\alpha/2].$$

□

Lemma 4. *If for all t the cumulative distribution functions of $R(\mathbf{S}_t)$ starting from $\mathbf{S}_0 = \mathbf{s}$ have finitely many jumps, then for all but finitely many points ν the function $\tilde{f}_{\mathbf{s}}$ is differentiable with derivative*

$$\frac{d}{d\nu} \tilde{f}_{\mathbf{s}}(\nu) = \sum_{k=1}^K \mathbb{E}_{\mathbf{s}} \left[h_{[k,1],n}^{(k)} \left(U_{[k,1],n}^{(k)}(\nu), \nu \right) \mathbb{I}(\tilde{V}_{\mathbf{s}}(\nu) \in [-1/2, 1/2]) \right],$$

where $\tilde{V}_{\mathbf{s}}(\nu) = \sum_{k=1}^K \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu)$, $U_{\mathbf{i},n}^{(k)}(\nu) = \arg \min_{u \in [N]} Z_{i,u,n}^{(k)}(\nu)$, and

$$\begin{aligned} h_{i,n}^{(1)}(t, \nu) &= (1 - \gamma^t) / (2(R_u - R_\ell)(1 - \gamma^N)), \\ h_{i,n}^{(k)}(t, \nu) &= h_{[i,1],n}^{(k-1)}(t, \nu) - \frac{1}{n(\mathbf{i}, k)} \sum_{j=2}^{n(\mathbf{i}, k) + 1} h_{[i,j],n}^{(k-1)}(U_{[i,j],n}^{(k-1)}(\nu), \nu) \Big| \{\mathbf{S}_{[0,t]}^{[i,j]} = \mathbf{S}_{[0,t]}^i\} \quad (k \geq 2). \end{aligned}$$

Proof. First, we show by induction that for all k, ν , random times T and sets E in $\times_i \sigma(\mathbf{S}^i)$ (the smallest product sigma algebra measuring all versions \mathbf{S}^i) we have

$$Z_{i,T,n}^{(k)}(\nu) \Big| E = A_{i,T,n}^{(k)}(\nu) \Big| E + \nu \cdot B_{i,T,n}^{(k)}(\nu) \Big| E$$

where $A_{i,t,n}^{(k)}, B_{i,t,n}^{(k)}$ are random variables only depending on ν through the minimizers $U_{i,n}^{(k')}(\nu)$ for $k' \leq k$.

For $k = 1$ the result is immediately verified for $A_{i,t,n}^{(1)}(\nu) = -\frac{(1-\gamma)}{(R_u - R_\ell)(1-\gamma^N)} \sum_{u=0}^{t-1} \gamma^u R(\mathbf{S}_u^i)$ and $B_{i,t,n}^{(1)}(\nu) = b_t$.

Now assume the result holds up to k , then by (8), as $\{\mathbf{S}_{[0,T]}^{[i,j]} = \mathbf{S}_{[0,T]}^{[i]}\} \in \times_i \sigma(\mathbf{S}^i)$ for all j

$$Z_{i,T,n}^{(k)}(\nu) \mid E = A_{i,T,n}^{(k)} \mid E + \nu \cdot B_{i,T,n}^{(k)} \mid E$$

for

$$A_{i,T,n}^{(k)} = A_{[i,1],T,n}^{(k-1)} - \frac{1}{n(i,k)} \sum_{j=2}^{n(i,k)+1} A_{[i,j],U_{[i,j],n}^{(k-1)}(\nu),n}^{(k-1)} \{\mathbf{S}_{[0,T]}^{[i,j]} = \mathbf{S}_{[0,T]}^{[i]}\},$$

$$B_{i,T,n}^{(k)} = B_{[i,1],T,n}^{(k-1)} - \frac{1}{n(i,k)} \sum_{j=2}^{n(i,k)+1} B_{[i,j],U_{[i,j],n}^{(k-1)}(\nu),n}^{(k-1)} \{\mathbf{S}_{[0,T]}^{[i,j]} = \mathbf{S}_{[0,T]}^{[i]}\}.$$

Second, we show that all but finitely many points ν we have

$$\mathbb{P}\left(\lim_{\nu' \rightarrow \nu} U_{i,n}^{(k)}(\nu') = U_{i,n}^{(k)}(\nu)\right) = 1.$$

For $k \geq 2$ and $t \in [N]^{\prod_{k'=1}^{k-1} (1+n(i,k'))}$ let $A_{i,t,n}^{(k)}$ be $A_{i,T,n}^{(k)}$ where every nested random time is replaced by the elements of \mathbf{t} consecutively, and let $b_{i,t,n}^{(k)}$ be defined similarly for $B_{i,T,n}^{(k)}$, note that this makes $b_{i,t,n}^{(k)}$ deterministic.

For $\nu_2 < \nu_1$

$$\begin{aligned} \mathbb{I}(U_{i,n}^{(k)}(\nu_1) \neq U_{i,n}^{(k)}(\nu_2)) &\leq \mathbb{I}(\exists t, u \ Z_{i,t,n}^{(k)}(\nu_1) \leq Z_{i,u,n}^{(k)}(\nu_1), Z_{i,t,n}^{(k)}(\nu_2) > Z_{i,u,n}^{(k)}(\nu_2)) \\ &\leq \mathbb{I}\left(\exists t, u, A_{i,t,n}^{(k)} + \nu_1 \cdot b_{i,t,n}^{(k)} \leq A_{i,u,n}^{(k)} + \nu_1 \cdot b_{i,u,n}^{(k)}, A_{i,t,n}^{(k)} + \nu_2 \cdot b_{i,t,n}^{(k)} > A_{i,u,n}^{(k)} + \nu_2 \cdot b_{i,u,n}^{(k)}\right) \\ &\leq \sum_{t,u} \mathbb{I}\left(A_{i,t,n}^{(k)} + \nu_1 \cdot b_{i,t,n}^{(k)} \leq A_{i,u,n}^{(k)} + \nu_1 \cdot b_{i,u,n}^{(k)}, A_{i,t,n}^{(k)} + \nu_2 \cdot b_{i,t,n}^{(k)} > A_{i,u,n}^{(k)} + \nu_2 \cdot b_{i,u,n}^{(k)}\right) \\ &= \sum_{t,u} \mathbb{I}\left(\frac{A_{i,t,n}^{(k)} - A_{i,u,n}^{(k)}}{b_{i,u,n}^{(k)} - b_{i,t,n}^{(k)}} \in (\nu_2, \nu_1]\right) \end{aligned}$$

The indicators above converge to $\mathbb{I}\left(\frac{A_{i,u,n}^{(k)} - A_{i,t,n}^{(k)}}{b_{i,t,n}^{(k)} - b_{i,u,n}^{(k)}} = \nu_1\right)$ when $\nu_2 \uparrow \nu_1$ which has nonzero probability of being equal to one for at most finitely many points ν_1 by the assumption. The direction $\nu_1 \downarrow \nu_2$ can be shown by changing the first bound above to

$$\mathbb{I}(\exists t, u \ Z_{i,t,n}^{(k)}(\nu_1) < Z_{i,u,n}^{(k)}(\nu_1), Z_{i,t,n}^{(k)}(\nu_2) \geq Z_{i,u,n}^{(k)}(\nu_2))$$

and following the same steps as above.

Third, we show for points ν_1 where all functions $U_{i,n}^{(k')}$ are continuous for $k' \leq k$ that

$$\frac{d}{d\nu} \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu) = h_{[k,1],n}^{(k)}(U_{[k,1],n}^{(k)}(\nu), \nu). \quad (54)$$

Note that, by the second step above, the complement of this set is a subset of the set \mathcal{V} of jump points for the cumulative distribution functions of $(A_{i,u,n}^{(k')} - A_{i,t,n}^{(k')})(b_{i,t,n}^{(k')} - b_{i,u,n}^{(k)})$ for all $i, t, u, k' \leq k$.

We show the result by induction. Let $k = 1$, then we have for $\nu_1 > \nu_2$

$$\begin{aligned} \frac{\min_{u \in [N]} Z_{[1,1],u,n}^{(1)}(\nu_1) - \min_{u \in [N]} Z_{[1,1],u,n}^{(1)}(\nu_2)}{\nu_1 - \nu_2} &\geq \frac{Z_{[1,1],U_{[1,1],n}^{(1)}(\nu_1),n}^{(1)}(\nu_1) - Z_{[1,1],U_{[1,1],n}^{(1)}(\nu_1),n}^{(1)}(\nu_2)}{\nu_1 - \nu_2} = b_{U_{[1,1],n}^{(1)}(\nu_1)}, \\ \frac{\min_{u \in [N]} Z_{[1,1],u,n}^{(1)}(\nu_1) - \min_{u \in [N]} Z_{[1,1],u,n}^{(1)}(\nu_2)}{\nu_1 - \nu_2} &\leq \frac{Z_{[1,1],U_{[1,1],n}^{(1)}(\nu_2),n}^{(1)}(\nu_1) - Z_{[1,1],U_{[1,1],n}^{(1)}(\nu_2),n}^{(1)}(\nu_2)}{\nu_1 - \nu_2} = b_{U_{[1,1],n}^{(1)}(\nu_2)}. \end{aligned}$$

By the choice of ν_1 we have that the upper and lower bound almost everywhere almost surely converge to $h_{[1,1],n}^{(1)}(U_{[1,1],n}^{(1)}(\nu_1), \nu_1) = b_{U_{[1,1],n}^{(1)}(\nu_1)}$ when $\nu_2 \uparrow \nu_1$. The case $\nu_1 < \nu_2$ can be shown similarly.

Let the above statement hold up to k , then for $\nu_1 > \nu_2$

$$\begin{aligned}
& \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_1) - \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_2) \\
&= Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_1),n}^{(k-1)}(\nu_1) - \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_1),n}^{(k-1)}(\nu_1) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1]} \right\} \right) \right. \\
&\quad \left. - Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_2),n}^{(k-1)}(\nu_2) + \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_2),n}^{(k-1)}(\nu_2) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_2) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_2) \rrbracket}^{[k,1]} \right\} \right) \right) \\
&\geq Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_1),n}^{(k-1)}(\nu_1) - \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_2),n}^{(k-1)}(\nu_1) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1]} \right\} \right) \right. \\
&\quad \left. - Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_1),n}^{(k-1)}(\nu_2) + \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_2),n}^{(k-1)}(\nu_2) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1]} \right\} \right) \right) \\
&= \left(Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_1),n}^{(k-1)}(\nu_1) - Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_1),n}^{(k-1)}(\nu_2) \right) \\
&\quad + \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_2),n}^{(k-1)}(\nu_1) - Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_2),n}^{(k-1)}(\nu_2) \right) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1]} \right\} \right).
\end{aligned}$$

Taking $\nu_2 \uparrow \nu_1$ above we see that

$$\liminf_{\nu_2 \uparrow \nu_1} \frac{\min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_1) - \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_2)}{\nu_1 - \nu_2} \geq h_{[k,1],n}^{(k)}(U_{[k,1],n}^{(k)}(\nu_1), \nu_1).$$

Similarly, it can be shown that

$$\begin{aligned}
& \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_1) - \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_2) \\
&\leq \left(Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_2),n}^{(k-1)}(\nu_1) - Z_{[k,1,1],U_{[k,1],n}^{(k)}(\nu_2),n}^{(k-1)}(\nu_2) \right) \\
&\quad + \frac{1}{n(\mathbf{i},k)} \sum_{j=2}^{n(\mathbf{i},k)+1} \left(Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_1),n}^{(k-1)}(\nu_1) - Z_{[k,1,j],U_{[k,1,j],n}^{(k-1)}(\nu_1),n}^{(k-1)}(\nu_2) \right) \left| \left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_2) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_2) \rrbracket}^{[k,1]} \right\} \right).
\end{aligned}$$

Now, by the choice of ν_1 , we know there is a random variable $\epsilon(\nu_1)$ such that $U_{[k,1],n}^{(k)}(\nu_2) = U_{[k,1],n}^{(k)}(\nu_1)$ almost surely for all ν_2 such that $|\nu_1 - \nu_2| \leq \epsilon(\nu_1)$. Hence, for every sample path, for ν_2 close enough to ν_1 , we condition on $\left\{ \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1,j]} = \mathbf{S}_{\llbracket 0,U_{[k,1],n}^{(k)}(\nu_1) \rrbracket}^{[k,1]} \right\}$ above. From this it follows that almost surely

$$\limsup_{\nu_2 \uparrow \nu_1} \frac{\min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_1) - \min_{u \in [N]} Z_{[k,1],u,n}^{(k)}(\nu_2)}{\nu_1 - \nu_2} \leq h_{[k,1],n}^{(k)}(U_{[k,1],n}^{(k)}(\nu_1), \nu_1)$$

hence (54) holds. The case $\nu_1 < \nu_2$ works similarly.

By induction it can be verified that the right-hand side of (54) is bounded for each k .

We have for $\nu_1 \in \mathcal{V}$ and $\nu_1 > \nu_2$

$$\begin{aligned}
\frac{\tilde{f}_s(\nu_1) - \tilde{f}_s(\nu_2)}{\nu_1 - \nu_2} &= \mathbb{E} \left[\sum_{k=1}^K \frac{\min_{u \in [N]} Z_{[1,j],u,n}^{(k)}(\nu_1) - \min_{u \in [N]} Z_{[1,j],u,n}^{(k)}(\nu_2)}{\nu_1 - \nu_2} \mathbb{I}((\tilde{V}_s(\nu_1), \tilde{V}_s(\nu_2)) \in [-1/2, 1/2]^2) \right] \\
&\quad + \mathbb{E} \left[\frac{V_s(\nu_1) - V_s(\nu_2)}{\nu_1 - \nu_2} \mathbb{I}((\tilde{V}_s(\nu_1), \tilde{V}_s(\nu_2)) \notin [-1/2, 1/2]^2) \right].
\end{aligned}$$

When $\nu_2 \uparrow \nu_1$ by continuity and boundedness of the derivative the first term goes to

$$\sum_{k=1}^K \mathbb{E}_{\mathbf{s}} \left[h_{[k,1],n}^{(k)} \left(U_{[k,1],n}^{(k)}(\nu), \nu \right) \mathbb{I}(\tilde{V}_{\mathbf{s}}(\nu) \in [-1/2, 1/2]) \right]$$

while the second term goes to zero by continuity (of $\tilde{V}_{\mathbf{s}}$) and boundedness, where in both cases the dominated convergence theorem was used. \square

Theorem 9. Let $h_{m,[k,1],n}^{(k)}$, $U_{m,[k,1],n}^{(k)}$ be independent versions (in m) of $h_{[k,1],n}^{(k)}$, $U_{[k,1],n}^{(k)}$ and

$$h_m : \nu \mapsto \sum_{k=1}^K h_{m,[k,1],n}^{(k)}(U_{m,[k,1],n}^{(k)}(\nu), \nu).$$

Let

$$\alpha_m = \frac{1}{|\sum_{\ell=1}^m h_{\ell}(\nu_{\ell}(\mathbf{s}))|}.$$

Let \mathcal{V} be the set of points where $\tilde{f}_{\mathbf{s}}$ is differentiable. If $\inf_{\nu \in \mathcal{V}} d/d\nu \tilde{f}_{\mathbf{s}}(\nu) > 0$, then there is a unique point $\tilde{\nu}(\mathbf{s})$ such that $\tilde{f}_{\mathbf{s}}(\tilde{\nu}(\mathbf{s})) = 0$. If the derivative of $f_{\mathbf{s},n}^{(K)}$ exists at $\tilde{\nu}(\mathbf{s})$ and $\mathbb{E}[V_{\mathbf{s}}(\tilde{\nu}(\mathbf{s}))^2] > 0$, we have

$$\sqrt{m}(\nu_m(\mathbf{s}) - \tilde{\nu}(\mathbf{s})) \xrightarrow{d} \mathcal{N} \left(0, \frac{\mathbb{E}[V_{\mathbf{s}}(\tilde{\nu}(\mathbf{s}))^2]}{\left(\frac{d}{d\nu} f_{\mathbf{s},n}^{(K)}(\tilde{\nu}(\mathbf{s})) \right)^2} \right). \quad (55)$$

Proof. The first statement is trivial as $\tilde{f}_{\mathbf{s}}$ has a positive derivative wherever the derivative is defined. We show the second statement by first verifying that the Robbins-Monro conditions hold almost surely, hence by Corollary 1 we have that $\nu_m(\mathbf{s})$ converges almost surely to $\tilde{\nu}(\mathbf{s})$. As for a constant $H < \infty$ we have $|h_{\ell}| \leq H$ we have that $\sum_{\ell=1}^m \alpha_{\ell} \geq \sum_{\ell=1}^m 1/(\ell H) \rightarrow \infty$ almost surely. We verify that $\sum_{\ell=1}^m \alpha_{\ell}^2 \leq \infty$ almost surely. Let $\xi_{\ell} = h_{\ell}(\nu_{\ell}(\mathbf{s})) - \mathbb{E}[h_1(\nu_{\ell}(\mathbf{s}))]$ and $\sigma_{\ell}^2 = \mathbb{E}[\xi_{\ell}^2 | \nu_{\ell}(\mathbf{s})]$. By the strong law of large numbers for martingales we have (by boundedness of ξ_{ℓ}) that almost surely

$$\lim_{m \rightarrow \infty} \frac{\sum_{\ell=1}^m \xi_{\ell}}{\sum_{\ell=1}^m \sigma_{\ell}^2} = 0 \quad (56)$$

Letting H_{σ} be such that $|\sigma_{\ell}^2| \leq H_{\sigma}$ for all ℓ we can choose $\epsilon \in (0, \inf_{\nu} \mathbb{E}[h_1(\nu)]/H_{\sigma})$ and have an $M \in \mathbb{N}$ such that for all $m > M$

$$\frac{1}{m} \sum_{\ell=1}^m h_{\ell}(\nu_{\ell}(\mathbf{s})) > \inf_{\nu} \mathbb{E}[h_1(\nu)] - \epsilon H_{\sigma} > 0.$$

We then conclude

$$\sum_{\ell=1}^{\infty} \alpha_{\ell}^2 \leq \sum_{\ell=1}^M \alpha_{\ell}^2 + \sum_{\ell=M+1}^{\infty} \frac{1}{m^2 \left(\frac{1}{m} \sum_{\ell=1}^m h_{\ell}(\nu_{\ell}(\mathbf{s})) \right)^2} < \sum_{\ell=1}^M \alpha_{\ell}^2 + \sum_{\ell=M+1}^{\infty} \frac{1}{m^2 (\inf_{\nu} \mathbb{E}[h_1(\nu)] - \epsilon H_{\sigma})^2} < \infty.$$

Hence it follows by Corollary 1 that $\nu_m(\mathbf{s}) \xrightarrow{a.s.} \tilde{\nu}(\mathbf{s})$.

Now, if $\mathbb{P}(h_{\ell}((\tilde{\nu}(\mathbf{s}))^-) \neq h_{\ell}((\tilde{\nu}(\mathbf{s}))^+)) > 0$ the derivative of $\tilde{f}_{\mathbf{s}}$ at $\tilde{\nu}$ would not exist, hence almost surely for all ℓ we have $h_{\ell}((\tilde{\nu}(\mathbf{s}))^-) = h_{\ell}((\tilde{\nu}(\mathbf{s}))^+)$ and combined with the above result it follows that $h_{\ell}(\nu_{\ell}(\mathbf{s})) \rightarrow h(\tilde{\nu}(\mathbf{s}))$ (with $h \stackrel{d}{=} h_1$ independently) almost surely by continuous mapping (noting that the set of discontinuity points of h is restricted to a deterministic set). By boundedness we also have $\mathbb{E}[h_{\ell}(\nu_{\ell}(\mathbf{s}))] \rightarrow \mathbb{E}[h(\tilde{\nu}(\mathbf{s}))]$. Similarly, it can be shown that $\sigma_{\ell}^2 \rightarrow \mathbb{E}[\xi_1^2 | \tilde{\nu}(\mathbf{s})] = \sigma^2$ which is deterministic. Hence $\frac{1}{m} \sum_{\ell=1}^m \sigma_{\ell}^2 \rightarrow \sigma^2$ and by (56) and continuous mapping we have

$$m\alpha_m = \frac{m}{|\sum_{\ell=1}^m h_{\ell}(\nu_{\ell}(\mathbf{s}))|} \rightarrow 1/\mathbb{E}[h_1(\tilde{\nu}(\mathbf{s}))].$$

The result follows along the lines of Lai and Robbins [1978], with three additional remarks.

- In order to use the representation (17) in Lai and Robbins [1978] the recursion of the stochastic approximation has to be truncated earlier, to make sure that the iterates $\nu_{\ell}(\mathbf{s})$ remain in an interval around $\tilde{\nu}(\mathbf{s})$ where $\tilde{f}_{\mathbf{s}}$ is differentiable, so as to use the mean value theorem.

- The martingale central limit theorem (Theorem 2 in Brown [1971]) can be used to show a central limit theorem result for the martingale $\sum_{\ell} \epsilon_{\ell}$, as the running mean of the quadratic variation process converges to a constant, we can show a central limit theorem result by just dividing by \sqrt{m} .
- A supremum law of iterated logarithm for martingales Fisher [1986] can be used to show that the middle term in (29) in Lai and Robbins [1979] goes to zero when divided by \sqrt{m} .

□

B Notation Table

Symbol	Definition	Defined in
a	Markov chain/arm	Sec. 2
t, u	Time index	Sec. 2
$\mathbf{s}_h^a, \mathbf{s}^a, \mathbf{s}$	Initial state of an arm	Sec. 2
$\mathbb{E}_{\mathbf{s}}, \mathbb{P}_{\mathbf{s}}$	Transition kernel and expectation for the Markov chains, conditional on initial state \mathbf{s}	Sec. 2
\mathcal{H}	Set of histories	(1)
R	Common reward function for the arms	Sec. 2
π, π^*	Policy and optimal policy (resp.) for the family of alternative bandit processes	Sec. 2
$C(\mathbf{s}^a)$	Total discounted absolute reward for sampling arm a , starting from state \mathbf{s}^a	Sec. 2
\mathbb{E}_{π}	Expectation operator under a fixed policy π	Sec. 2
γ	Discount factor	Sec. 2
$\nu(\mathbf{s})$	Gittins index for state \mathbf{s}	Sec. 2
$\mathcal{F}_t^a, \mathcal{F}_t$	Natural filtration generated by \mathcal{S}^a and \mathcal{S} (resp.), including starting state	Sec. 2, 3.1
$\mathcal{T}^a, \mathcal{T}$	Set of stopping times w.r.t. $(\mathcal{F}_t^a)_t$ or $(\mathcal{F}_t)_t$ (resp.)	Sec. 2, 3.1
$\mathcal{S}_t^a, \mathcal{S}_t$	State of Markov chain/arm a and general arm (resp.) at time t	Sec. 2, 3.1
τ	Stopping time, used to determine optimal stopping value	Sec. 2
N	Sampling horizon for optimal stopping value	Sec. 3.1
\mathcal{T}_N	Set of stopping times adapted to \mathcal{F}_t and bounded by N	Sec. 3.1
g_t	Measurable real-valued cost function	Sec. 3.1
$Z_t, Z_t^{(1)}$	Cost $g_t(\mathcal{S}_{[0,t]})$ of an arm up to time t	Sec. 3.1
$Z_t^{(k+1)}$	$Z_t^{(k)} - \mathbb{E}[\min_{u \in [N]} Z_u^{(k)} \mid \mathcal{F}_t]$	Sec. 3.1
K	Truncation point of number of nested expectations in optimal stopping approximation	Sec. 3.1
$Z_{i,t,n}^{(k+1)}$	Sampling-based approximation of $Z_t^{(k+1)}$	Sec. 3.2
$n(i, k)$	Number of simulated paths used to determine $Z_{i,t,n}^{(k+1)}$	Sec. 3.2
$V_n^{(K)}, V_s(\nu), V_{s,m}(\nu)$	Sampling-based approximation of $\sum_{k=1}^K \mathbb{E}[\min_{u \in [N]} Z_u^{(k)}]$ and $\sum_{k=1}^K \mathbb{E}_s[\min_{u \in [N]} Z_u^{(k)}(\nu)]$	Sec. 3.2, 4
σ	Stopping time, used to truncate the support of the rewards	(15)
R_u, R_{ℓ}	Upper and lower bound for reward support (resp.), induced by stopping time σ	Sec. 4
c	Constant, equal to $(1 - \gamma)/(2(R_u - R_{\ell})(1 - \gamma^N))$	Sec. 4
$Z_t(\nu)$	Cost $c \sum_{u=0}^{t \wedge \sigma - 1} \gamma^u (\nu - R(\mathcal{S}_u))$ of an arm up to the minimum of time t and σ	Sec. 4
$\nu_{\sigma}(\mathbf{s})$	Gittins index approximation found by truncation of horizon and reward support	(16)
$\nu_m(\mathbf{s})$	Stochastic approximation iterates for determining SBGIA	(18)
α_m	Step-size sequence	Sec. 4
$\nu_M(\mathbf{s})$	Sampling-based Gittins index approximation	Sec. 4
$\tilde{f}_s(\nu)$	Expectation of $\mathbb{E}_s[V_s(\nu)]$	(19)
$f_s(\nu)$	Optimal stopping value $\inf_{\tau \in \mathcal{T}_N} \mathbb{E}_s[Z_{\tau}(\nu)]$	(19)
$B(\delta, \xi)$	Domain error bound, equal to $\xi + \delta$	(20)
ϵ_m	Martingale difference sequence	(21)
$\bar{\nu}_m, \underline{\nu}_m$	Stochastic approximation iterates bounding ν_m from above and below (resp.)	(22), (23)
\mathcal{R}_s	Set of roots of \tilde{f}_s	Sec. 4.2
$\tilde{\nu}(\mathbf{s})$	Root of \tilde{f}_s	Sec. 4.2
h_m	Derivative of $V_{s,m}(\nu)$	Th. 9
Ψ_t	Sufficient statistics	Sec. 5.2.1
κ_t	Effective number of observations	Sec. 5.2.1
O_t^a	Outcome t for distribution a	Sec. 5.1
θ_a	Parameter for distribution a	Sec. 5.1
$\Pi_t^a, \hat{\Pi}_t^a$	Posterior and approximate posterior (resp.) for θ_a based on $(O_u^a)_{u=1}^t$	Sec. 5.1, (51)