Decomposition-Coordination Method for Finite Horizon Bandit Problems

Michel de Lara, Benjamin Heymann, Jean-Philippe Chancelier

Optimality Gap Analysis of the Decomposition-Coordination Method for Finite Horizon Bandit Problems

Benjamin Heymann*, Michel De Lara†, Jean-Philippe Chancelier*

May 7, 2024

Abstract

This study explores multi-armed bandit problems under the premise that the decision-maker possesses prior knowledge of the arms' distributions and knows the finite time horizon. These conditions render the problems suitable for stochastic multistage optimization decomposition techniques. On the one hand, multi-armed bandit algorithms are integral to reinforcement learning and are supported by extensive theoretical analysis, particularly regarding regret minimization. Meanwhile, stochastic multistage optimization emphasizes examining the performance gap between a given policy and the optimal, adapted policy. Within this framework, decomposition strategies have been demonstrated to efficiently tackle large-scale stochastic multistage optimization challenges. Our empirical findings corroborate the approach's efficacy on multi-armed bandit. Most importantly, we contribute a visual interpretation of the optimality gap between relaxed and admissible solutions, which lays the groundwork for subsequent investigations into the performance of decomposition methods.

Keywords. multi-armed bandit problem, dynamic programming, decomposition-coordination, Lagrangian relaxation

1 Introduction

A multi-armed bandit (MAB) is a mathematical model for sequential decision making under partial feedback. At each round, a decision maker selects an arm, and then the arm yields a random reward that depends on the intrinsic characteristics of the arm, which are unknown to the decision maker. The selection of an arm hence serves two purposes: amassing reward and acquiring information on the arm, to be used in the future rounds. For this reason, MABs embody the well-known exploration-exploitation tradeoff and have concentrated the attention of the research community for decades.

The first occurrence of MABs in the literature was motivated by clinical trials [31], but the rise of the digital economy has unlocked many new applications [12, 27, 33]. As observed in [10], two schools emerged from the early works on MABs. The first school follows [4] and aims at maximizing the expected total reward over a discounted horizon, and envisions the multi-armed bandit as a Markov Decision Process (MDP). The pioneering breakthrough is the Gittins Index Theorem [17] which provides a way to decompose the problem into tractable subproblems, one per arm. The second school follows Robbins formulation [29] and seeks to minimize an expected regret over a finite horizon. The seminal work [24, 23] identifies an asymptotically efficient policy. Other problem formulations and approaches were proposed following this milestone [26, 7, 2].

In this article, we take the MDP perspective, and aim at maximizing the intertemporal expected reward or, equivalently, minimizing the Bayesian regret of a binary bandit over a known, finite timespan. Theoretically, such a problem could be addressed with dynamic programming, but this is not feasible because of the curse of dimensionality: it is well known that the problem size grows exponentially in the number of arms.

*Criteo Technology, Paris, France
†CERMICS, Ecole des Ponts, Marne-la-Vallée, France

We leverage the ideas from [19, 1, 8, 6] to show how time decomposition (dynamic programming) can be made compatible with arm decomposition (Lagrangian relaxation). This results in a time-dependent index policy.

Our approach connects to the literature on Lagrangian relaxation of weakly coupled stochastic dynamic programs [34, 19, 1, 6]. Our coupling constraint corresponds to the premise that only one arm can be pulled at any stage.

Asymptotic results for such type of algorithms exist for restless bandits [35, 6], when the number of arms that can be pulled is a constant factor of the time horizon. However, as far as we know, no study has been done on MAB with finite horizon. We focus, like in [16], on small time horizons.

We follow [8], where it is shown how a structured, large scale intertemporal maximization problem can be transformed into a collection of parameterized, simpler, intertemporal subproblems by relaxing coupling constraints. Thus doing, one obtains a collection of local value functions, one per subproblems, all functions of a common coordinating parameter process. After optimizing this latter, one sums the local value functions and uses the resulting surrogate global value function in the online phase of the Bellman equation. This gives both a theoretical upper bound for the original maximization problem, and a heuristic online policy.

This article presents a numerical test bench for the use of the weakly coupled constraint decomposition method (for short, DeCo) applied to the Bayesian bandits problem, and presents an analysis of the optimality gap. The decomposition procedure is not new, and it has been in particular used recently in [6] for a setting close to ours (dynamic selection). Experimentally, what differentiate us from [6] is a focus on a different problem, namely, the Bayesian binary bandit, because (a) this is a very active area of research for which many solutions have been proposed, (b) it is a "simpler" problem for which we derive theoretical insights on why those decomposition methods work so well in practice. The literature on decomposition of weakly coupled problems does not tell much on the optimality gap in nonasymptotic setting. As such, the characterisation of the optimality gap presented in this paper, while focused on a specific problem, is a step toward better understanding this class of methods. While it does not fully explain why DeCo works so well, it is a step into that direction as it provides a visual interpretation of the gap between the upper bound provided by the relaxed solution and the performance of DeCo.

The paper is organized as follows. In Sect. 2, we describe the stochastic multi-armed bandit problem, and we show how it can be treated by decomposition-coordination using a methodology that we refer to as DeCo . In Sect. 3, we present an interpretation of DeCo 's decision rule in term of value of information. In Sect. 4 we provide a geometric interpretation of the optimality gap between relaxed solutions and admissible solutions.We present some algorithmic aspects of DeCo in Sect. 5. In Sect. 6, we show numerically that DeCo achieves state-of-the-art performance. It is remarkable that, on our numerical experiments, DeCo offers performances comparable to [30], while also showing empirically good performances for both small number of arms and for the "many arms" regime.

2 Preliminaries

In §2.1, we describe the stochastic multi-armed bandit problem, and show how it can be framed in the multistage stochastic optimal control formalism. In §2.2, we present the arm decompsition method in [6, 8], and we show how this control problem can be treated by decomposition-coordination.

2.1 The Bayesian multi-armed bandit problem

We now present the (binary) Bayesian multi-armed bandit problem using the formalism of multistage stochastic optimal control. For any integers $r \leq s$, $[\![r, s]\!]$ denotes the subset $\{r, r+1, \ldots, s-1, s\}$. We consider a decision-maker (DM) who selects an arm a in a finite set A, at each discrete time stage t in the set $[\![0, T-1]\!]$, where $T \geq 1$ is an integer, the *horizon*. Thus doing, the arm a delivers, at the end of the time interval $[t, t+1[$,

a random variable[1] $\mathbf{W}^a_{t+1}$ that takes two values[2] in the set $\{\mathtt{B},\mathtt{G}\}$ ("bad" $\mathtt{B}$, "good" $\mathtt{G}$) and with unknown parameter $\bar{p}^a_{\mathtt{G}}$, the probability of the event $\mathbf{W}^a_{t+1}=\mathtt{G}$. The parameter $\bar{p}^a_{\mathtt{G}}\in[0,1]$ is unknown to the DM, which we formalize below.

Probabilistic model Let $\Sigma=\big\{p=(p^{\mathtt{B}},p^{\mathtt{G}})\in\mathbb{R}^2\,\big|\,p^{\mathtt{B}}\geq 0\ ,\ \ p^{\mathtt{G}}\geq 0\ ,\ \ p^{\mathtt{B}}+p^{\mathtt{G}}=1\big\}$ be the one-dimensional simplex[3]. For any $p=(p^{\mathtt{B}},p^{\mathtt{G}})\in\Sigma$, we denote by $\mathcal{B}(p^{\mathtt{B}},p^{\mathtt{G}})=\bigotimes_{t=1}^{T}\big(p^{\mathtt{B}}\delta_{\mathtt{B}}+p^{\mathtt{G}}\delta_{\mathtt{G}}\big)$, where δ denotes a Dirac measure, the probability on $\{\mathtt{B},\mathtt{G}\}^T$ which corresponds to a sequence of independent (Bernoulli) random variables with values in $\{\mathtt{B},\mathtt{G}\}$. For $\{p^a\}_{a\in A}=\{(p^{\mathtt{B}a},p^{\mathtt{G}a})\}_{a\in A}\in\prod_{a\in A}\Sigma$, we consider the probability $\bigotimes_{a\in A}\mathcal{B}(p^{\mathtt{B}a},p^{\mathtt{G}a})$ on the product space $\prod_{a\in A}\{\mathtt{B},\mathtt{G}\}^T$, which corresponds to independence between arms in A. We denote by $\mathbb{E}_{\{p^a\}_{a\in A}}$ the corresponding mathematical expectation. Let $\Delta(\Sigma)$ denote the set of probability distributions on the simplex Σ. We suppose that the DM holds a prior $\pi^a_0\in\Delta(\Sigma)$ over the unknown $p^a=(p^{\mathtt{B}a},p^{\mathtt{G}a})\in\Sigma$, for every arm $a\in A$. In practice, we will consider a beta distribution $\beta(n^{\mathtt{B}},n^{\mathtt{G}})$ on Σ, with positive integers $n^{\mathtt{B}}>0$ and $n^{\mathtt{G}}>0$ as parameters.

We consider the probability space $(\Omega,\mathcal{F},\mathbb{P})$ where $\Omega=\prod_{a\in A}\Sigma\times\{\mathtt{B},\mathtt{G}\}^T$, $\mathcal{F}=\bigotimes_{a\in A}\mathcal{B}^{\mathrm{o}}_{\Sigma}\otimes 2^{(\{\mathtt{B},\mathtt{G}\}^T)}$ (where $\mathcal{B}^{\mathrm{o}}_{\Sigma}$ is the Borel σ-field over the simplex Σ), $\mathbb{P}=\bigotimes_{a\in A}\pi^a_0\big(\mathrm{d}(p^{\mathtt{B}a},p^{\mathtt{G}a})\big)\otimes\mathcal{B}(p^{\mathtt{B}a},p^{\mathtt{G}a})$. Then, $\mathbf{W}^a=\{\mathbf{W}^a_t\}_{t\in[\![1,T]\!]}$ denotes the coordinate mappings for every arm $a\in A$, with $\mathbf{W}^a_t$ a random variable having values in the set $\{\mathtt{B},\mathtt{G}\}$. For a given family $\{(\bar{p}^a_{\mathtt{B}},\bar{p}^a_{\mathtt{G}})\}_{a\in A}\in\prod_{a\in A}\Sigma$ and for $\pi^a_0=\delta_{(\bar{p}^a_{\mathtt{B}},\bar{p}^a_{\mathtt{G}})}$, for every arm $a\in A$, the family $\{\mathbf{W}^a_t\}_{a\in A,t\in[\![1,T]\!]}$ consists of independent random variables, where $\mathbf{W}^a_t$ has (Bernouilli) probability distribution with parameter $\bar{p}^a_{\mathtt{G}}\in[0,1]$, that is, $\mathbb{P}\big(\mathbf{W}^a_t=\mathtt{B}\big)=1-\bar{p}^a_{\mathtt{G}}$ and $\mathbb{P}\big(\mathbf{W}^a_t=\mathtt{G}\big)=\bar{p}^a_{\mathtt{G}}$. With this probabilistic model, we represent the sequential independent outcomes of $|A|$ independent arms (where $|A|$ denotes the cardinality of the finite set A).

Decision model We consider a sequence $\mathbf{U}=\{\mathbf{U}_t\}_{t\in[\![0,T-1]\!]}$ of random variables (on the probability space $(\Omega,\mathcal{F},\mathbb{P})$), where $\mathbf{U}_t=\{\mathbf{U}^a_t\}_{a\in A}$, $\mathbf{U}^a_t\in\{0,1\}$, $\forall a\in A$, $\forall t\in[\![0,T-1]\!]$. Their possible values in $\{0,1\}$ represent that either arm a has been selected at the beginning of the time interval $[t,t+1[$ ($\mathbf{U}^a_t=1$) or not ($\mathbf{U}^a_t=0$). Since, at each given time t, one and only one arm has to be selected, we add the (almost sure) constraint

$$\sum_{a\in A}\mathbf{U}^a_t=1\ ,\ \ \forall t\in[\![0,T-1]\!]\ . \tag{1}$$

Such way of modelling the selection of a fixed number of arms dates back to the restless bandit paper [34], where Whittle replaces the almost sure constraint by an expectation. It has been applied for similar problems in [10, 6].

Information and admissible controls When the arm a has been selected at stage t (that is, when $\mathbf{U}^a_t=1$), the DM observes the outcome, in the set $\{\mathtt{B},\mathtt{G}\}$, of the random variable $\mathbf{W}^a_{t+1}$. When the arm a has not been selected at stage t (that is, when $\mathbf{U}^a_t=0$), the DM observes nothing (zero 0). Thus, the DM observes the random variable $\mathbf{Y}_{t+1}=\big\{\mathbf{U}^a_t\mathbf{W}^a_{t+1}\big\}_{a\in A}$, which takes values from the set $\{\mathtt{B},\mathtt{G},0\}$ for all $t\in[\![0,T-1]\!]$. Then, the admissible controls $\mathbf{U}=\{\mathbf{U}_t\}_{t\in[\![0,T-1]\!]}$ are those that satisfy

$$\sigma(\mathbf{U}_0)=\{\emptyset,\Omega\}\ \text{and} \tag{2}$$

$$\sigma(\mathbf{U}_t)\subset\sigma(\mathbf{U}_0,\mathbf{Y}_1,\mathbf{U}_1\ldots,\mathbf{U}_{t-1},\mathbf{Y}_t)\ ,\ \ \forall t\in[\![1,T-1]\!]\ , \tag{3}$$

where $\sigma(\mathbf{Z})\subset\mathcal{F}$ is the σ-field generated by the random variable $\mathbf{Z}$ on the probability space $(\Omega,\mathcal{F},\mathbb{P})$.

[1]The shifted index $t+1$ is here to indicate that the random variable $\mathbf{W}^a_{t+1}$ materializes during the time interval $[t,t+1[$.

[2]We call these two values "bad" (for $\mathtt{B}$), and "good" (for $\mathtt{G}$), and not $\{0,1\}$ to avoid confusion with the possible values for the controls ("do not select arm", "select arm"). In fact, we take two values for the sake of simplicity, but we could have taken a finite or even infinite number of values.

[3]For the sake of symmetry between outcomes $\mathtt{B}$ and $\mathtt{G}$, we do not identify the simplex Σ with the unit segment $[0,1]$ by the mapping $\Sigma\ni(p^{\mathtt{B}},p^{\mathtt{G}})\mapsto p^{\mathtt{B}}\in[0,1]$.

Random rewards We consider given a family $\{L_t^a\}_{a\in A, t\in[\![0,T-1]\!]}$ of functions $L_t^a : \{\mathtt{B},\mathtt{G}\} \to \mathbb{R}$, that represent instantaneous rewards as follows. When the arm a has been selected at stage t (that is, when $\mathbf{U}_t^a = 1$), the random variable $\mathbf{W}_{t+1}^a$ materializes and the DM receives the payoff $1 \times L_t^a(\mathbf{W}_{t+1}^a) = \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$. When the arm a has not been selected at stage t (that is, when $\mathbf{U}_t^a = 0$), the DM receives the payoff $0 = \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$. Thus, the total random reward associated with the *control sequence* $\mathbf{U} = \{\mathbf{U}_t\}_{t\in[\![0,T-1]\!]}$ is given by $\sum_{t=0}^{T-1}\sum_{a\in A}\mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$.

State variable For any arm $a \in A$ and control sequence $\mathbf{U} = \{\mathbf{U}_t\}_{t\in[\![0,T-1]\!]}$, we set

$$\mathbf{N}_t^{a,\mathtt{G}} = \sum_{s=0}^{t-1} \mathbf{U}_t^a \mathbf{1}_{\{\mathbf{W}_{t+1}^a = \mathtt{G}\}} \text{ and } \mathbf{N}_t^{a,\mathtt{B}} = \sum_{s=0}^{t-1} \mathbf{U}_t^a \mathbf{1}_{\{\mathbf{W}_{t+1}^a = \mathtt{B}\}} .$$

$\mathbf{N}_t^{a,\mathtt{G}}$ (resp. $\mathbf{N}_t^{a,\mathtt{B}}$) represents the quantities of good (resps. bad) pulls that the decision maker has observed up to time $t-1$ (right before making a decision at time t) when pulling arm a, When two different control sequences are envisioned, it is helpful to be able to explicite the dependency of $\mathbf{N}$ on $\mathbf{U}$, so that we also introduce the notation

$$\mathbf{N}_t^{\mathbf{U},a} = \left(\mathbf{N}_t^{a,\mathtt{G}} \ , \ \mathbf{N}_t^{a,\mathtt{B}}\right) . \tag{4}$$

Knowing $\mathbf{N}_t^{\mathbf{U},a}$, the expected reward of pulling arm a can be deduced by the DM through posterior update. This justifies the introduction of the following functions $\ell_t^a : \mathbb{N}^2 \to \mathbb{R}$ given by

$$\forall (n^{\mathtt{B}a}, n^{\mathtt{G}a}) \in \mathbb{N}^2 \ , \ \ell_t^a(n^{\mathtt{B}a}, n^{\mathtt{G}a}) = \frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} L_t^a(\mathtt{B}) + \frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} L_t^a(\mathtt{G}) , \tag{5}$$

which corresponds to the arm expected reward computed according to the updated posterior.

Optimality criteria in the Bayesian framework Let $\Delta(\Sigma)$ denote the set of probability distributions on the simplex Σ. We denote by $\pi_0 = \{\pi_0^a\}_{a\in A} \in \prod_{a\in A}\Delta(\Sigma)$ the family of initial priors, one for each arm, and we formulate the following maximization problem — where the supremum is taken over $\mathbf{U} = \{\mathbf{U}_t^a\}_{a\in A, t\in[\![0,T-1]\!]} \in \{0,1\}^{A\times[\![0,T-1]\!]}$, subject to constraints (1) and (3),

$$V_0(\pi_0) = \sup \int_{\Delta(\Sigma)^{|A|}} \prod_{a\in A} \pi_0^a(\,\mathrm{d}p^a) \mathbb{E}_{\{p^a\}_{a\in A}}\left[\sum_{t=0}^{T-1}\sum_{a\in A} \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)\right] \tag{6a}$$

$$\text{s.t.} \quad \sum_{a\in A} \mathbf{U}_t^a = 1 \ , \ \ \forall t \in [\![0,T-1]\!] \tag{6b}$$

$$\sigma(\mathbf{U}_t) \subset \sigma(\mathbf{U}_0, \mathbf{Y}_1, \ldots, \mathbf{U}_{t-1}, \mathbf{Y}_t) \ , \ \ \forall t \in [\![1,T-1]\!] \ , \ \ \sigma(\mathbf{U}_0) = \{\emptyset, \Omega\} . \tag{6c}$$

We denote by $\mathcal{U}^{ad}$ the set of controls satisfying the constraints (1) and (3) (or, equivalently, constraints (6b) and (6c)) of Problem (6). For any control sequence $\mathbf{U} \in \mathcal{U}^{ad}$, we set the *intertemporal expected reward* or *total reward*

$$V_0^{\mathbf{U}}(\pi_0) = \int_{\Delta(\Sigma)^{|A|}} \prod_{a\in A} \pi_0^a(\,\mathrm{d}p^a) \mathbb{E}_{\{p^a\}_{a\in A}}\left[\sum_{t=0}^{T-1}\sum_{a\in A} \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)\right] . \tag{7}$$

2.2 Arm Decomposition

The stochastic optimal control problem (6) is, theoretically, solvable by dynamic programming, but the fact that the computing cost grows exponentially fast with the problem size prevents us from doing so in

practice. One trick leveraged by several authors [6, 8] consist in relaxing the set of constraints (6b). This is done in two steps. First the almost sure constraints (6b) are replaced by constraints in expectation: $\mathbb{E}\big[\sum_{a\in A}\mathbf{U}_t^a\big]=1$. Second, the constraints in expectation are replaced by a penalty term in the criterion $-\mu_t\big(\mathbb{E}\big[\sum_{a\in A}\mathbf{U}_t^a\big]-1\big)$, where the vector of penalty parameters $(\mu_t)_{t\in[\![0,T-1]\!]}$ serve as a lever to push the system to satisfy the constraint. Those two steps result in the following modified problem:

$$V_0^r[\mu](\pi_0)=\sup\int_{\Delta(\Sigma)^{|A|}}\prod_{a\in A}\pi_0^a(\,\mathrm{d}p^a)\mathbb{E}_{\{p^a\}_{a\in A}}\Big[\sum_{t=0}^{T-1}\sum_{a\in A}\mathbf{U}_t^a\left(L_t^a(\mathbf{W}_{t+1}^a)-\mu_t\right)\Big]+\sum_{t=0}^{T-1}\mu_t \tag{8a}$$

$$\text{s.t.}\quad \sigma(\mathbf{U}_t)\subset\sigma(\mathbf{U}_0,\mathbf{Y}_1,\dots,\mathbf{U}_{t-1},\mathbf{Y}_t)\ ,\ \ \forall t\in[\![1,T-1]\!]\ ,\ \ \sigma(\mathbf{U}_0)=\{\emptyset,\Omega\}\ . \tag{8b}$$

We refer to $V_t^r[\mu]$ as the *value of the relaxed problem*. Because of their Lagrangian interpretation, the vector of penalty parameter $(\mu_t)_{t\in[\![0,T-1]\!]}$ is often referred to as the vector of multipliers. The resulting new problem has the crucial property of being equivalent to $|A|$ independent, smaller subproblems. As opposed to the initial problem (6a), those subproblems can be addressed with dynamic programming. More precisely, the value of the relaxed problem writes

$$V_0^r[\mu]\big((n^{\mathtt{B}a})_{a\in A},(n^{\mathtt{G}a})_{a\in A}\big)= \tag{9}$$

$$\sum_{a\in A}V_0^a[\mu](n_0^{\mathtt{B}a},n_0^{\mathtt{G}a})+\sum_{t=0}^{T-1}\mu_t, \tag{10}$$

where $V_0^a[\mu](n_0^{\mathtt{B}a},n_0^{\mathtt{G}a})$ is the value of the subproblem associated with the backward induction:

$$V_T^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a})=0\ ,$$

$$\begin{aligned}V_t^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a})&=\max\Big\{V_{t+1}^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a}),-\mu_t+\frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}}\big(L_t^a(\mathtt{B})+V_{t+1}^a[\mu](n^{\mathtt{B}a}+1,n^{\mathtt{G}a})\big)\\&\qquad+\frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}}\big(L_t^a(\mathtt{G})+V_{t+1}^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a}+1)\big)\Big\}\end{aligned} \tag{11a}$$

$$=\max\Big\{V_{t+1}^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a}),-\mu_t+\ell_t^a(n^{\mathtt{B}a},n^{\mathtt{G}a})+\overline{\overline{V_{t+1}^a[\mu]}}(n^{\mathtt{B}a},n^{\mathtt{G}a})\Big\}\ , \tag{11b}$$

and where in the last line, we used the notation

$$\forall(n^{\mathtt{B}},n^{\mathtt{G}})\in\mathbb{N}^2\ ,\ \ \bar{\varphi}(n^{\mathtt{B}},n^{\mathtt{G}})=\frac{n^{\mathtt{B}}}{n^{\mathtt{B}}+n^{\mathtt{G}}}\varphi(n^{\mathtt{B}}+1,n^{\mathtt{G}})+\frac{n^{\mathtt{G}}}{n^{\mathtt{B}}+n^{\mathtt{G}}}\varphi(n^{\mathtt{B}},n^{\mathtt{G}}+1)\ . \tag{12}$$

It is well known [6, 8] that $V_0^r[\mu]$ is an upper-bound on the value of the stochastic optimal control problem (6) V_0, which is what we express in the next proposition.

Proposition 1 *We have the upper bound*

$$V_0\big((n_0^{\mathtt{B}a})_{a\in A},(n_0^{\mathtt{G}a})_{a\in A}\big)\le\inf_{\mu\in\mathbb{R}^T}\Big(V_0^r[\mu]\Big)\ . \tag{13}$$

where we identify (by an abuse of notation) $V_0\big((n_0^{\mathtt{B}a})_{a\in A},(n_0^{\mathtt{G}a})_{a\in A}\big)$ *with the value* $V_0(\pi_0)$ *of problem* (6) *when the prior* $\pi_0=\{\beta(n_0^{\mathtt{B}a},n_0^{\mathtt{G}a})\}_{a\in A}$ *, is a Beta distribution.*

In what follows we propose to generate a policy using $V^r[\mu]$ instead of the "true" value function in Bellman equation, for a suitable choice of μ. We call this procedure DeCo . This procedure is not new, and has been in particular used recently in [6] for a setting close to ours (dynamic selection).

It is left as an exercise to check that, when the state of the multi-armed system is given by $n=(n_t^{\mathtt{B}a},n_t^{\mathtt{G}a})_{a\in A}\in\prod_{a\in A}\mathbb{N}\times\mathbb{N}$ at time t, the DeCo algorithm selects an arm that maximizes the value-to-go, or otherwise said, selects an arm $a^\star$ in[4]

$$\forall n\in\mathbb{N}^{2|A|}\ ,\ \ A^\sharp(t,\mu,n)=\arg\max_{a\in A}\big[-V_{t+1}^a[\mu](n^{\mathtt{B}a},n^{\mathtt{G}a})+\ell_t^a(n^{\mathtt{B}a},n^{\mathtt{G}a})+\overline{\overline{V_{t+1}^a[\mu]}}(n^{\mathtt{B}a},n^{\mathtt{G}a})\big]\ . \tag{14}$$

[4]In case of non uniqueness, take any arm in the arg max.

3 The value of information interpretation

Next we provide an interpretation of the online decision rule of DeCo stated in (14). If we set

$$\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) = \overline{\overline{V_{t+1}^a[\mu]}}(n^{\mathtt{B}a}, n^{\mathtt{G}a}) - V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) \,, \tag{15}$$

then we can interpret $\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})$ as the "value of information" in the decomposed subproblem of arm a when the vector of multipliers is μ: it is the incremental performance an optimal policy gets if given an additional pull outcome. The higher $\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})$, the higher a pull of a increase the expected value one can get from a in the later rounds in the decomposed problem. Hence, $\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})$ quantify the value of exploration. It is insightful to observe that the selected arm during the DeCo online phase (see Equation (14)) is the one that maximizes

$$\underbrace{I_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{index}} = \underbrace{\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{value of information (exploration)}} + \underbrace{\ell_t^a(n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{reward (exploitation)}} \,. \tag{16}$$

We recognize an exploration and an exploitation term. Such exploration term is reminiscent of the exploration term encountered in the Upper Confidence Bound (UCB) algorithms [2]. Also, [18] refers to a learning component in the Gittins index as the difference between the index value and the immediate expected reward. More recently, the notion of information gain is also important in [30].

In particular, the definition (11) of the individual Bellman values in Proposition 1 becomes

$$\begin{aligned} \underbrace{V_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{current Bellman value}} &= \max\left\{ V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}), -\mu_t + \ell_t^a(n^{\mathtt{B}a}, n^{\mathtt{G}a}) + \overline{\overline{V_{t+1}^a[\mu]}}(n^{\mathtt{B}a}, n^{\mathtt{G}a}) \right\} , && \text{(by (11b))} \\ &= \underbrace{V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{future Bellman value at the same state}} + \underbrace{\left(I_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) - \mu_t \right)^+}_{\text{incremental gain from pulling}} , \end{aligned} \tag{17}$$

where $x^+ = \max(x, 0)$, which means that the arm is pulled in the decomposed problem only if the sum (16) of the information gain (δ_t^a) and the expected reward (ℓ_t^a) is greater than μ_t. Hence μ_t can be interpreted as an equilibrium price of a "bandit market". In this bandit market, each bandit is handled by an independent profit maximizing agent, and the agent is required to pay the market price μ_t to pull the arm of her/his bandit at time t. This is different but connected to the fair charge metaphore proposed in [32] for the Gittins index. The important nuance is that, here, the price depends on a market made of several arms whereas, for the Gittins index, the fair charge is arm specific.

4 Characterization of the optimality gap

In the absence of the constraint (1) of pulling only one arm, the solutions of the subproblems (one per arm) could be aggregated into an admissible solution of the original problem (6). From this perspective, the aggregation of the subproblems solutions constitutes a solution to the relaxed problem (8).

This section introduces, in §4.1, a geometric interpretation of the gap between a relaxed solution and any admissible solution (Theorem 2). We then specialize this result to DeCo (Theorem 6) in §4.2. The idea is then illustrated with simulation in § 4.3.

4.1 Comparative analysis of admissible and relaxed solution performances

We first show a general result that allows us to compare the total reward (7) generated by an admissible solution and the total reward generated by a (possibly non admissible) solution to the relaxed problem (8).

For $\mu = (\mu_0, \dots, \mu_{T-1}) \in \mathbb{R}_+^T$, $a \in A$, $t \in [0 \dots T-1]$, we define

$$\forall (n^{\mathtt{B}a}, n^{\mathtt{G}a}) \in \mathbb{N}^2 \,, \;\; u_t^{a,r}[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) = \begin{cases} 1 & \text{if} \quad I_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) - \mu_t > 0 \,, \\ 0 & \text{elsewhere,} \end{cases} \tag{18}$$

which can be interpreted as a one-arm optimal policy associated with $V_t^a[\mu](n^a)$, where $n^a = (n^{\mathtt{B}a}, n^{\mathtt{G}a})$. The following result compares the quantities $V_0^{\mathbf{U}}$ (defined in (7)) and $V_0^r[\mu]$(defined in (8)). Proposition 1 ensures that the former is smaller than the latter. The next result, Theorem 2, quantifies the gap.

Theorem 2 *Let $\mu = (\mu_0, \dots, \mu_{T-1}) \in \mathbb{R}_+^T$, and $\mathbf{U} \in \mathcal{U}^{ad}$, and $n \in \mathbb{N}^{2|A|}$. Then, we have that*

$$V_0^r[\mu](n) - V_0^{\mathbf{U}}(n) = \mathbb{E}_n\Big[\sum_{t=0}^{T-1}\sum_{a\in A}\Big(I_t^a[\mu](\mathbf{N}_t^{\mathbf{U},a}) - \mu_t\Big)\cdot \underbrace{\Big(u_t^{a,r}[\mu](\mathbf{N}_t^{\mathbf{U},a}) - \mathbf{U}_t^a\Big)}_{\in\{0,-1,+1\}}\Big]\,, \tag{19}$$

where the value function $V_0^r[\mu]$ is defined in Equation (9) and the value function $V_0^{\mathbf{U}}(n)$ is defined by

$$V_0^{\mathbf{U}}(n) = \mathbb{E}_n\Big[\sum_{t=0}^{T-1}\sum_{a\in A}\ell_t^a(\mathbf{N}_t^{\mathbf{U}})\cdot\mathbf{U}_t^a(\mathbf{N}_t^{\mathbf{U}})\Big]\,, \tag{20}$$

and $\mathbb{E}_n[\cdot]$ denotes an expectation where the underlying state process takes value n at time 0, that is $\mathbf{N}_0^{\mathbf{U}} = n$.

Proof. We fix a strategy $\mathbf{U} : \{n^a\}_{a\in A} \to \{0,1\}^{|A|}$. We start by a two preliminary facts.

• Using postponed Lemma 3, we obtain that the function $V_0^{\mathbf{U}}$ defined in Equation (20) coincides with the function $V_0^{\ell,\mathbf{U}}$ where the sequence of value functions $(V_t^{\ell,\mathbf{U}})_{t\in[\![0,T]\!]}$ satisfy the Bellman equation (25), that is

$$V_t^{\ell,\mathbf{U}}(n) = \mathcal{B}_{t+1}^{\ell,\mathbf{U}}[V_{t+1}^{\ell,\mathbf{U}}](n) = \ell_t\big(n,\mathbf{U}_t(n)\big) + \sum_{a\in A}\overline{\overline{V_{t+1}^{\mathbf{U}}(n^{(-a)},\cdot)}}(n^a)\cdot\mathbf{U}_t^a(n)\,, \tag{21}$$

where the sequence of mappings $(\ell_t)_{t\in[\![0,T-1]\!]}$ is defined, for all $n \in \mathbb{N}^{2|A|}$ and all $v \in \{0,1\}^{|A|}$, by $\ell_t(n,v) = \sum_{a\in A}\ell_t^a(n^a)\cdot v^a$.

• Using again the postponed Lemma 3, we obtain that the right hand side of Equation 19 is equal to the value function $\Gamma_0^{\mathbf{U}}(n)$ where the sequence of value functions $(\Gamma_t^{\mathbf{U}})_{t\in[\![0,T]\!]}$ is solution of the Bellman equation

$$\Gamma_T^{\mathbf{U}} \equiv 0 \text{ and } \forall t\in[\![0,T-1]\!]\,,\ \Gamma_t^{\mathbf{U}} = \mathcal{B}_{t+1}^{\gamma,\mathbf{U}}[\Gamma_{t+1}^{\mathbf{U}}]\,, \tag{22a}$$

where

$$\forall\varphi : \mathbb{N}^{2|A|}\to\overline{\mathbb{R}}\,,\ \forall n\in\mathbb{N}^{2|A|}\,,\ \mathcal{B}_{t+1}^{\gamma,\mathbf{U}}[\varphi](n) = \gamma_t\big(n,\mathbf{U}_t(n)\big) + \sum_{a\in A}\Big(\overline{\overline{\varphi(n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}_t^a(n)\,. \tag{22b}$$

and where the sequence of functions $(\gamma_t)_{t\in[\![0,T-1]\!]}$ is defined by

$$\forall n\in\mathbb{N}^{2|A|}\,,\ \forall v\in\{0,1\}^{|A|}\,,\ \gamma_t(n,v) = \sum_{a\in A}\big(I_t^a[\mu](n)-\mu_t\big)\cdot\big(u_t^{a,r}[\mu](n)-v^a\big)\,. \tag{22c}$$

Finally, to prove Equation 19, we prove that the sequence of functions $(V_t^r[\mu] - V_t^{\mathbf{U}})_{t\in[\![0,T]\!]}$, where the functions $(V_t^r[\mu])_{t\in[\![0,T-1]\!]}$ are defined by

$$\forall t\in[\![0,T]\!]\,,\ V_t^r[\mu](n) = \sum_{s=t}^{T-1}\mu_s + \sum_{a\in A}V_t^a[\mu](n^a)\,, \tag{23}$$

satisfy the Bellman equation (22) — Note that at time 0, Equation (23) coincides with Equation (9) defining function $V_0^r[\mu]$ —.

First, for $t = T$ we immediately check that $V_T^r[\mu] - V_T^{\mathbf{U}} = 0$. Second, for $t \in [\![0,T-1]\!]$ and $n \in \mathbb{N}^{2|A|}$, we successively have

$$\mathcal{B}^{\gamma,\mathbf{U}}_{t+1}\big[V^r_{t+1}[\mu]-V^{\mathbf{U}}_{t+1}\big](n)=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{a\in A}\Big(\overline{\overline{(V^r_{t+1}[\mu]-V^{\mathbf{U}}_{t+1})(n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}^a_t(n) \qquad \text{(by (22b))}$$

$$=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{a\in A}\Big(\overline{\overline{V^r_{t+1}[\mu](n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}^a_t(n)-\sum_{a\in A}\Big(\overline{\overline{V^{\mathbf{U}}_{t+1})(n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}^a_t(n)$$

$$=\mathcal{B}^{\gamma,\mathbf{U}}_{t+1}\big[V^r_{t+1}[\mu]\big](n)-\sum_{a\in A}\Big(\overline{\overline{V^{\mathbf{U}}_{t+1})(n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}^a_t(n) \qquad \text{(by (22b))}$$

$$=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{s=t+1}^{T-1}\mu_s+\sum_{a\in A}\mathcal{T}^{\mathbf{U},a}\big[V^a_{t+1}[\mu]\big](n^a)-\sum_{a\in A}\Big(\overline{\overline{V^{\mathbf{U}}_{t+1})(n^{(-a)},\cdot)}}(n^a)\Big)\cdot\mathbf{U}^a_t(n)\,,$$

(by Lemma 4 applied to separable function V^r_{t+1} given by (23))

$$=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{s=t+1}^{T-1}\mu_s+\sum_{a\in A}\mathcal{T}^{\mathbf{U},a}\big[V^a_{t+1}[\mu]\big](n^a)+\sum_{a\in A}\ell^a_t(n^a)\cdot\mathbf{U}^a_t(n)-V^{\mathbf{U}}_t(n) \qquad \text{(by (21))}$$

$$=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{s=t+1}^{T-1}\mu_s+\sum_{a\in A}\Big(V^a_{t+1}[\mu](n^a)+\big(\overline{\overline{V^a_{t+1}[\mu]}}(n^a)-V^a_{t+1}[\mu](n^a)+\ell^a_t(n^a)\big)\cdot\mathbf{U}^a_t(n)\Big)-V^{\mathbf{U}}_t(n)$$

(by definition of $\mathcal{T}^{U,a}$ in (27))

$$=\gamma_t\big(n,\mathbf{U}_t(n)\big)+\sum_{s=t+1}^{T-1}\mu_s+\sum_{a\in A}\Big(V^a_{t+1}[\mu](n^a)+I^a_t[\mu](n^a)\cdot\mathbf{U}^a_t(n)\Big)-V^{\mathbf{U}}_t(n) \qquad \text{(by (15)-(16))}$$

$$=\sum_{a\in A}\big(I^a_t[\mu](n^a)-\mu_t\big)\cdot\big(u^{a,r}_t[\mu](n^a)-\mathbf{U}^a_t(n)\big)+\sum_{s=t+1}^{T-1}\mu_s$$
$$+\sum_{a\in A}\Big(V^a_{t+1}[\mu](n^a)+I^a_t[\mu](n^a)\cdot\mathbf{U}^a_t(n)\Big)-V^{\mathbf{U}}_t(n) \qquad \text{(by definition of } \gamma_t \text{ in (22c))}$$

$$=\sum_{a\in A}\Big(\big(I^a_t[\mu](n^a)-\mu_t\big)\cdot u^{a,r}_t[\mu](n^a)+V^a_{t+1}[\mu](n^a)\Big)+\sum_{s=t+1}^{T-1}\mu_s+\underbrace{\sum_{a\in A}\mu_t\cdot\mathbf{U}^a_t(n)}_{=\mu_t\ \text{as}\ \sum\mathbf{U}^a_t=1}-V^{\mathbf{U}}_t(n)$$

$$=\sum_{a\in A}\Big(\big(I^a_t[\mu](n^a)-\mu_t\big)^+V^a_{t+1}[\mu](n^a)\Big)+\sum_{s=t}^{T-1}\mu_s-V^{\mathbf{U}}_t(n) \qquad \text{(by definition of } u^{a,r}_t \text{ in (18))}$$

$$=\sum_{a\in A}V^a_t[\mu](n^a)+\sum_{s=t}^{T-1}\mu_s-V^{\mathbf{U}}_t(n) \qquad \text{(by (17))}$$

$$=V^r_t[\mu](n)-V^{\mathbf{U}}_t(n)\,. \qquad \text{(by definition of } V^r_t[\mu] \text{ in (23))}$$

This ends the proof. □

The following Lemmata are instrumental in the proof of Theorem 2.

Lemma 3 *Let be given $\mathbf{U}\in\mathcal{U}^{ad}$ and a sequence of function $(h_t)_{t\in[\![0,T-1]\!]}$ where $h_t:\mathbb{N}^{2|A|}\times\{0,1\}^{|A|}\to\overline{\mathbb{R}}$. Then, the sequence of value functions $(V^{h,\mathbf{U}}_t)_{t\in[\![0,T]\!]}$ defined for $t=T$ by $V^{h,\mathbf{U}}_T=0$ and for all $t\in[\![0,T-1]\!]$ by*

$$V^{h,\mathbf{U}}_t(n)=\mathbb{E}_n\Big[\sum_{s=t}^{T-1}h_s\big(\mathbf{N}^{\mathbf{U}}_s,\mathbf{U}_s(\mathbf{N}^{\mathbf{U}}_s)\big)\Big]\,, \tag{24}$$

satisfies the Bellman equation

$$V^{h,\mathbf{U}}_T\equiv 0\ \textit{and}\ \forall t\in[\![0,T-1]\!]\,,\ \ V^{h,\mathbf{U}}_t=\mathcal{B}^{h,\mathbf{U}}_{t+1}[V^{h,\mathbf{U}}_{t+1}]\,, \tag{25a}$$

where

$$\forall \varphi : \mathbb{N}^{2|A|} \to \overline{\mathbb{R}}\ ,\ \ \forall n \in \mathbb{N}^{2|A|}\ ,\ \ \mathcal{B}^{h,\mathbf{U}}_{t+1}[\varphi](n) = h_t\big(n, \mathbf{U}_t(n)\big) + \sum_{a\in A} \Big(\overline{\overline{\varphi(n^{(-a)},\cdot)}}(n^a)\Big) \cdot \mathbf{U}^a_t(n)\ . \tag{25b}$$

Proof. Left to the sagacity of the reader. □

Lemma 4 *Let $\mathbf{U} \in \mathcal{U}^{ad}$ be given. Consider a separable function $\varphi : \mathbb{N}^{2|A|} \to \overline{\mathbb{R}}$ defined by $\varphi(n) = \alpha + \sum_{a\in A} \varphi^a(n^a)$ then, for all $t \in [\![0, T-1]\!]$, we have that*

$$\forall n \in \mathbb{N}^{2|A|}\ ,\ \ \mathcal{B}^{h,\mathbf{U}}_{t+1}[\varphi](n) = h_t\big(n, \mathbf{U}_t(n)\big) + \alpha + \sum_{a\in A} \mathcal{T}^{\mathbf{U},a}[\varphi^a](n^a)\ , \tag{26}$$

where the Bellman operator is defined by Equation (25b) *in Lemma 3 and the notation $\mathcal{T}^{U,a}$ is defined as follows*

$$\forall \varphi : \mathbb{N}^2 \to \overline{\mathbb{R}}\ ,\ \ \forall n \in \mathbb{N}^{2|A|}\ ,\ \ \mathcal{T}^{\mathbf{U},a}[\varphi](n) = \varphi(n^a)\big(1 - \mathbf{U}^a(n)\big) + \overline{\overline{\varphi}}(n^a) \cdot \mathbf{U}^a(n)\ . \tag{27}$$

Proof. We consider a separable function φ given by $\varphi(n) = \alpha + \sum_{a\in A} \varphi^a(n^a)$ and we successively compute

$$\begin{aligned}
\mathcal{B}^{h,\mathbf{U}}_{t+1}[\varphi](n) &= h_t\big(n, \mathbf{U}_t(n)\big) + \sum_{a\in A} \Big(\overline{\overline{\varphi\big(n^{(-a)},\cdot\big)}}(n^a)\Big) \cdot \mathbf{U}^a_t(n^a) && \text{(by (25b))}\\
&= h_t\big(n, \mathbf{U}_t(n)\big) + \sum_{a\in A} \Big(\alpha + \sum_{a'\in A, a'\neq a} \varphi^{a'}(n^{a'}) + \overline{\overline{\varphi^a}}(n^a)\Big) \cdot \mathbf{U}^a_t(n^a) && \text{(as } \varphi \text{ is separable)}\\
&= h_t\big(n, \mathbf{U}_t(n)\big) + \alpha \underbrace{\sum_{a\in A} \mathbf{U}^a_t(n^a)}_{=1 \text{ as } \mathbf{U}\in\mathcal{U}^{ad}} + \sum_{a\in A} \sum_{a'\in A, a'\neq a} \varphi^{a'}(n^{a'}) \cdot \mathbf{U}^a_t(n^a) + \sum_{a\in A} \overline{\overline{\varphi^a}}(n^a) \cdot \mathbf{U}^a_t(n^a)\\
&= h_t\big(n, \mathbf{U}_t(n)\big) + \alpha + \sum_{a\in A} \varphi^a(n^a) \cdot \big(1 - \mathbf{U}^a_t(n^a)\big) + \sum_{a\in A} \overline{\overline{\varphi^a}}(n^a) \cdot \mathbf{U}^a_t(n^a) && \text{(by Lemma 5)}\\
&= h_t\big(n, \mathbf{U}_t(n)\big) + \alpha + \sum_{a\in A} \mathcal{T}^{\mathbf{U},a}[\varphi^a](n^a)\ . && \text{(by (27))}
\end{aligned}$$

This ends the proof. □

Lemma 5 *Let $g : A \to \mathbb{R}$ and $j : A \to \mathbb{R}$ be given and assume that $\sum_{a\in A} j(a) = 1$. Then, we have that*

$$\sum_{a\in A} \sum_{a'\in A, a'\neq a} g(a')j(a) = \sum_{a\in A} g(a) \cdot \big(1 - j(a)\big)\ . \tag{28}$$

Proof. We successively have

$$\begin{aligned}
\sum_{a\in A} \sum_{a'\in A, a'\neq a} g(a')j(a) &= \sum_{a\in A} \sum_{a'\in A} g(a')j(a)\mathbf{1}_{a\neq a'} = \sum_{a'\in A} g(a')\Big(\sum_{a\in A} j(a)\mathbf{1}_{a\neq a'}\Big)\\
&= \sum_{a'\in A} g(a')\Big(\underbrace{\sum_{a\in A} j(a)}_{=1} - j(a')\Big) = \sum_{a'\in A} g(a')\big(1 - j(a')\big)\ .
\end{aligned}$$

This ends the proof. □

4.2 Optimality gap estimate for DeCo

Next we specialize Theorem 2 to DeCo. That is, we consider Theorem 2 with the admissible policy used by DeCo denoted by DeCo$[\mu]$ defined for all $n = \{n^a\}_{a\in A} \in \mathbb{N}^{2|A|}$ by (see Equation 14)

$$\textsc{DeCo}[\mu]^a(n) = \begin{cases} 1 & \text{if } a = a^\star(n)\ , \\ 0 & \text{if } a \neq a^\star(n)\ , \end{cases} \tag{29}$$

where $a^\star(n)$ is a unique arm selected in $\arg\max_{a\in A} I_t^a[\mu](n)$.

Theorem 6 *Let $\mu = (\mu_0, \dots, \mu_{T-1}) \in \mathbb{R}_+^T$, and $n \in \mathbb{N}^{2|A|}$. Then, we have that*

$$V_0^r[\mu](n) - V_0^{\textsc{DeCo}[\mu]}(n) = \sum_{t\in[T]} \mathbb{E}\left[\left(\mu_t - I_t^{a^\star(\mathbf{N}_t^{\textsc{DeCo}})}[\mu](\mathbf{N}_t^{\textsc{DeCo}}) \right)^+ + \sum_{a\in A\left(t,\mu,\mathbf{N}_t^{\textsc{DeCo},a}\right)} \left(I_t^a[\mu](\mathbf{N}_t^{\textsc{DeCo},a}) - \mu_t \right) \right], \tag{30}$$

where $A(t,\mu,n) = \{a \,|\, a \neq a^\star(n) \wedge I_t^a[\mu](n) \geq \mu_t\}$; that is the set of arms with index $I_t^a[\mu]$ greater than μ_t which are not selected by the DeCo *policy.*

Proof. We first observe that, by definition of $u^{a,r}$ (see (18)), we have that for all $n = \{n^a\}_{a\in A} \in \mathbb{N}^{2|A|}$

$$\begin{aligned} &\left(I_t^a[\mu](n^a) - \mu_t \right) \cdot \left(u_t^{a,r}[\mu](n^a) - \mathbf{U}_t^a(n) \right) \\ &\qquad = \left(I_t^a[\mu](n^a) - \mu_t \right)^+ \cdot \left(1 - \mathbf{U}_t^a(n) \right) + \left(I_t^a[\mu](n^a) - \mu_t \right)^- \cdot \mathbf{U}_t^a(n)\ . \end{aligned} \tag{31}$$

where $x^- = \max(-x, 0)$. Now, using the notation $\psi = \textsc{DeCo}[\mu]$, we successively obtain that

$$\begin{aligned} V_0^r[\mu](n) - V_0^\psi(n) &= \mathbb{E}\left[\sum_{t\in[T]} \sum_{a\in A} \left(I_t^a[\mu](\mathbf{N}_t^\psi) - \mu_t \right) \cdot \left(u_t^{a,r}[\mu](\mathbf{N}_t^\psi) - \psi_t^a[\mu](\mathbf{N}_t^\psi) \right) \right] && \text{(by (19))} \\ &= \mathbb{E}\left[\sum_{t\in[T]} \sum_{a\in A} \left(I_t^a[\mu](\mathbf{N}_t^\psi) - \mu_t \right)^+ \cdot \left(1 - \psi_t^a(\mathbf{N}_t^\psi) \right) + \left(I_t^a[\mu](\mathbf{N}_t^\psi) - \mu_t \right)^- \cdot \psi_t^a(\mathbf{N}_t^\psi) \right] && \text{(by (31))} \\ &= \mathbb{E}\left[\sum_{t\in[T]} \left(\left(I_t^{a^\star(\mathbf{N}_t^\psi)}[\mu](\mathbf{N}_t^\psi) - \mu_t \right)^- + \sum_{a\in A, a\neq a^\star(\mathbf{N}_t^\psi)} \left(I_t^a[\mu](\mathbf{N}_t^\psi) - \mu_t \right)^+ \right) \right], && \text{(by (29))} \end{aligned}$$

which immediately gives Equation (30) and ends the proof. □

Since $V_0^r[\mu](n)$ is an upper bound for the performance of any admissible control, Theorem 6 indicates that the optimality gap is upper bounded by the sum of two terms. One term corresponds to the situations encountered by DeCo, when no arms has an index greater than μ_t. The other term corresponds to the situations when strictly more than one arm has an index greater than μ_t.

4.3 Illustration

We illustrate Theorem 6 with a numerical simulation reproduced four times. We use DeCo to produce an approximation of the optimal Bayesian control when 8 arms are uniformly sampled from $[0, 1]$, then we sample uniformly the values of 8 arms in $[0, 1]$, and run a simulation. We show the results in Figure 2. The optimality loss can be "read" directly from the plot.

5 Algorithmic aspects of DeCo

The DeCo algorithm is made of an offline computation phase (described in §5.1) and of an online computation phase (described in §5.2) phases as follows. The offline phase is summarized in Figure 1: it consists in the minimization of a dual function φ, where each evaluation of φ relies on solving $|A|$ independent Bellman equations. The minimization in (13) can be performed by gradient descent. Then, the online phase consists in using the upper bound function (13) as a proxy for the Bellman value, while ensuring that only one arm is pulled at each step. In §5.3, we discuss the computation cost of DeCo .

5.1 Offline phase of the DeCo algorithm

The offline phase of the DeCo algorithm is the minimization of the upper bound $V_0^r[\mu]$ with respect to μ (see (13)), for a family $\pi_0 = \{\pi_0^a\}_{a\in A} = \{\beta(n_0^{\mathtt{B}a}, n_0^{\mathtt{G}a})\}_{a\in A} = \beta(n_0)$ of beta priors. The algorithm is summarized

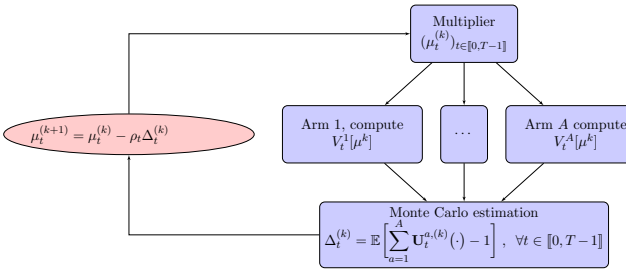

Figure 1: The decomposition coordination algorithm (DeCo)

in Figure 1 and its four steps are as follows.

(S_1) Choose an initial vector $\mu^{(0)} \in \mathbb{R}^T$ of multipliers.

(S_2) At iteration k, given a vector $\mu^{(k)} \in \mathbb{R}^T$ of multipliers, compute the collection $\left\{V_t^a[\mu^{(k)}]\right\}_{t\in[\![0,T]\!],a\in A}$. The computation is performed in parallel, arm per arm. Note that $V_t^a[\mu^{(k)}]$, the Bellman value function at time $t \in [\![0, T]\!]$, is to be evaluated only on the finite grid $\{(n_0^{\mathtt{B}a} + n^{\mathtt{B}a}, n_0^{\mathtt{G}a} + n^{\mathtt{G}a}) \,|\, n^{\mathtt{B}a} + n^{\mathtt{G}a} \leq t\}$. Note also that, when all the arms share the same prior and the same instantaneous reward, a unique sequence of Bellman value functions is to be computed, that is, all the arms share the same sequence of Bellman value functions.

(S_3) Once gotten the collection of $\left\{V_0^a[\mu^{(k)}]\right\}_{a\in A}$ of decomposed Bellman value functions at iteration k, update the vector of multipliers by a gradient step to obtain $\mu^{(k+1)}$. The gradient of the dual function φ with respect to the multipliers is obtained by computing the expectation of the dualized constraint as formulated in Problem (13) (see [9] for more details). Numerically, the expectation is obtained by Monte Carlo simulations. [5] In some of our numerical experiments, we use a solver (limited memory Bfgs) of the Modulopt library from Inria [15]. To obtain a global $O(T^3)$ running time, the computing budget allocated to this iterated gradient phase does not depend on T.

(S_4) Stop the iterations (stopping criterion) or go back to step S_2 with multiplier $\mu^{(k+1)}$.

5.2 Online phase of the DeCo algorithm

In the offline phase, we obtain a multiplier μ and a collection of value function V^a. During the online phase, the arms are selected according to the rule specified in Equation (14).

[5] The gradient phase to minimize (13) can be replaced by a more sophisticated algorithm such as the conjugate gradient or the quasi-Newton method.

The structure of policy (14) is that of a *nonstationary*[6] *index policy*. Indeed, the right hand side in (14) is a quantity that depends only on t and on the state $(n_t^{\mathtt{B}a}, n_t^{\mathtt{G}a})$ of arm a at time t. The DeCo policy used in numerical experiments is the policy $\mathcal{A}^*[\mu^*]$ in (14), where μ^* is given by the offline phase (described in §5.1) of the DeCo algorithm.

5.3 Computational complexity

Solving the maximization problem (6), that is, computing $V_0(\pi_0)$ for a given prior (like, for instance, the uniform distribution given by the beta distribution $\beta(1,1)$ for all arms) can be done using Dynamic programming. This is however, only possible for relatively small instances of problem (6), that is, for a limited number $|A|$ of arms and a limited time horizon T. We recall here that solving the problem for $|A|$ arms requires solving a Bellman equation with a state of dimension $2|A|$ (a state described by two integers per arm), which implies an exponential increase $O((2|A|)^T)$ in computational cost with respect to $|A|$. This is an instance of what Richard Bellman referred to as the *curse of dimensionality*. For FH-Gittins , [28] provides methodologies to compute FH-Gittins indices that are $O(T^6)$ in time complexity.

By contrast, for DeCo, each dynamic programming phase costs $O(T^3)$ in running time: indeed for each time $t = 1, \ldots, T$, we need a grid of $T \times T$ for the 2 dimensional prior parameter that counts the number of successes and failures, hence $O(T^3)$ in total, hence the complexity is $O(KT^3)$, where K is the number of arms with different parameters.

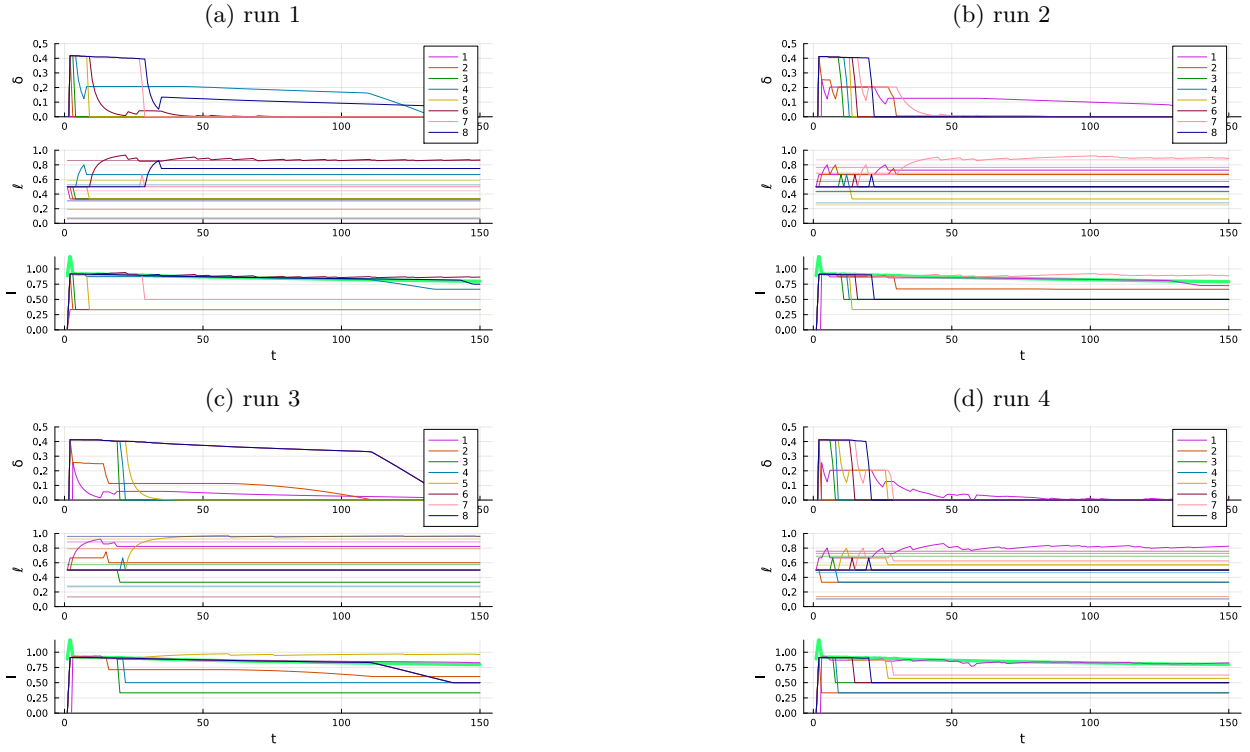

Figure 2: Four simulations of trajectories generated with the controls obtained from DeCo for a uniform prior and 8 arms. In each of the four subfigures, we display the value of information δ (top), the expected reward ℓ (middle) and the index I (bottom) obtained with DeCo. The green line corresponds to the value of the component μ_t of the multiplier at the end of the dual gradient procedure.

[6] If we had considered an infinite horizon, we would have obtained a (stationary) index policy.

6 Numerical experiments

In this section, we report numerical experiments for some short horizons instances of the Bayesian Bernoulli MAB problem. For the sake of reproducibility, we have performed two separate implementations with two different languages, one in Julia [5], the other in Nsp [11]. On the instances we have tested, DeCo competes with state-of-the-art policies for MABs.

First, in Table 1, we compare the performance of DeCo against the brute force approach Bf and FH-Gittins, (because of the computation cost of Bf, FH-Gittins is used as a proxy supposed to be close to the optimal solution). We observe that the performance of DeCo is close[7] to the optimal solution while keeping the computational cost reasonable (at most 1.3 second).

We then tested DeCo against Thomson Sampling (Ts) [31, 12], Kullback-Leibler upper-confidence bound (Kl-Ucb) [14], Information-Directed Sampling[8] (IDS) [30], Finite Horizon Gittins index FH-Gittins [22, 28, 25] and, in the case of two arms, the exact dynamic programming solution[9]. Any control sequence $\mathbf{U} = \{\mathbf{U}_t^a\}_{a\in A, t\in[\![0,T-1]\!]}$ is assessed (and compared to others) using the expected Bayesian regret given by

$$\mathcal{R}(\mathbf{U}) = \int_{\Delta(\Sigma)^A} \prod_{a\in A} \pi_0^a(\,\mathrm{d}p^a) \left\{ \mathbb{E}_{\{p^a\}_{a\in A}} \left[\sum_{t=0}^{T-1} \sum_{a\in A} \left(\mathbf{U}_t^{\mathrm{Ba},a} - \mathbf{U}_t^a\right) \mathbf{W}_{t+1}^a \right] \right\}, \tag{32}$$

where we have set the instantaneous costs L_t^a equal to 1 on `G` and 0 on `B` for and where the Ba (best arm) policy is, for all $a \in A$, given by $\mathbf{U}_t^{\mathrm{Ba},a} = 1 \iff a \in \arg\max_{a'\in A} p_{\mathtt{G}}^{a'}$, and where the prior is supposed to be the uniform distribution for all arms. Numerically, the expected Bayesian regret is obtained by Monte Carlo simulations, where the expectation with respect to the prior is obtained with a sample of size 1000 and expectation with respect to the arms parameters is obtained with a sample of size 1000 in Figure 3 and of size 100 in Figure 4. On all cases, DeCo beats both Ts and Kl-Ucb with a comfortable margin, and is comparable to IDS. For the two arms case in Figure 3(a), DeCo is very close to the optimal solution, computed by dynamic programming (we used the Julia BinaryBandit library [21, 20]).

Last, we also computed the dual bound provided by DeCo. Indeed, the upper bound (13) yields the inequality

$$\mathcal{R}(\mathbf{U}) \geq \mathcal{R}^{\mathrm{LB}} = \frac{|A|}{|A|+1} T - \left(\sum_{a\in A} V_0^a[\mu^*](\pi_0^a) + \sum_{t=0}^{T-1} \mu_t^* \right). \tag{33}$$

Figure 4 shows the regret lower bound, and the DeCo, Ts and Kl-Ucb regrets as a function of the number of arms for horizons $T = 100$ and $T = 500$ (beware: in Figure 4, the x axis is the number of arms!). The lower bound is of no use (lower than 0) for small numbers 2 and 5 of arms. Nevertheless, when the number of arms increases, the regret of DeCo and the lower bound become quite close, which indicates that, for those examples, DeCo is close to being optimal.

7 Conclusion

It is notable that decomposition methods perform so well out of the box, even when compared to the best known approaches to such a very scrutinized problem as is the Bayesian bandit. The numerical results demonstrate the value of the decomposition-coordination approach and shall serve as a proof of concept: DeCo is a simple algorithm and its performances are close to the optimal Bayesian solution for several configurations of arms and horizons, while keeping the computing time reasonable. Empirically, DeCo offer performances comparable to FH-Gittins but with a much smaller computation burden. For practical

[7] It might happen that DeCo empirical average is better than Bf, but this is due to the simulation noise.

[8] For IDS we used the library [3]

[9] We did not include Optimistic Gittins Index as we did not manage to reproduce the results in [13]. Also, it seems that the right hand side in [13, example 3.1] is inexact, as reveals the case $\lambda = 1$.

Table 1: Comparison of DeCo, Bf and FH-Gittins in term of estimated total expected reward (higher is better).

$\|A\|$	T	Bf	DeCo
3	10	6.409	6.411
3	20	13.465	13.458
5	10	6.659	6.645

$\|A\|$	T	FH-Gittins	DeCo
5	20	14.28	14.21
5	40	30.06	29.85
15	20	14.67	14.59
15	40	31.63	31.54

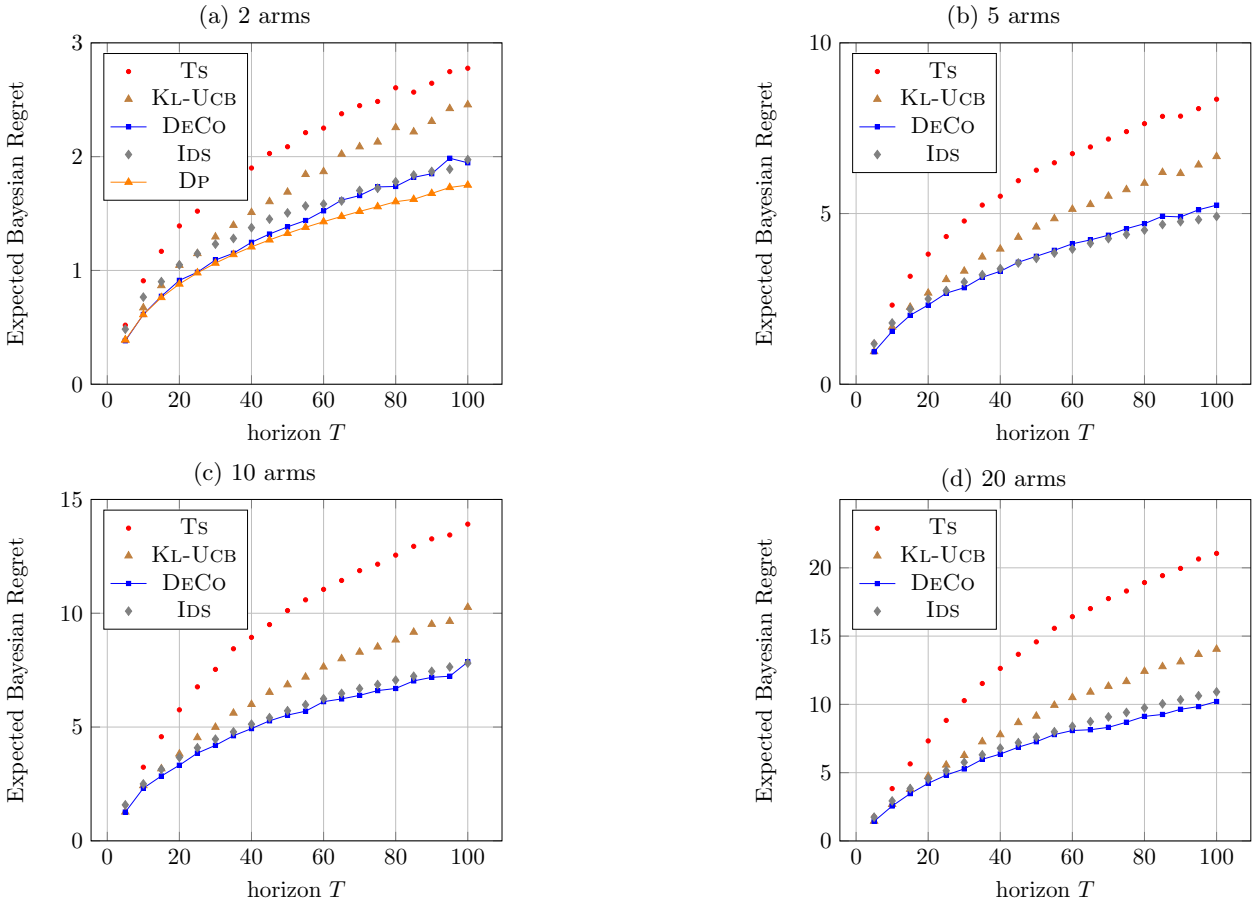

Figure 3: Expected Bayesian regret (32) for DeCo and a few benchmark policies (the lower the better) for 2, 5, 15 and 20 arms with uniform prior.

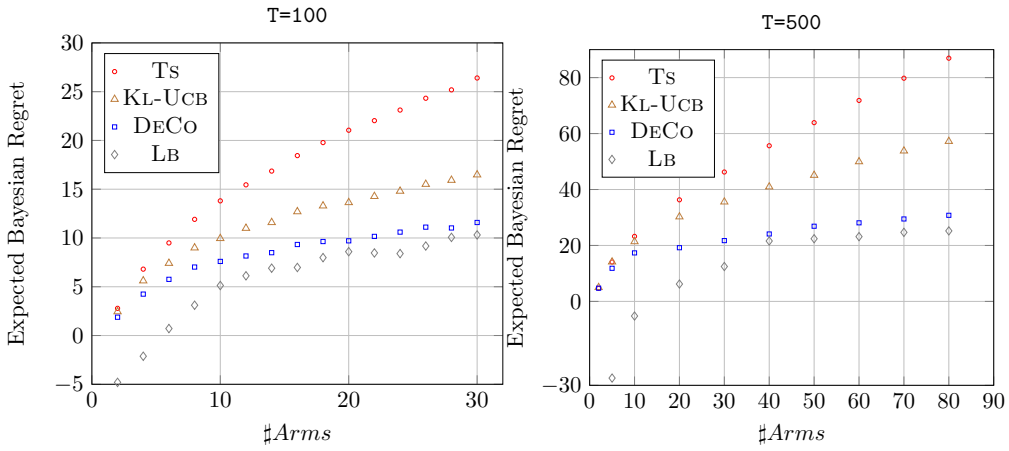

Figure 4: Expected Bayesian regret (32) for DeCo, Ts and Kl-Ucb with uniform prior, as functions of the number $\sharp Arms$ of arms. The (DeCo) lower bound Lb in (33) is also plotted and demonstrates that DeCo is close to the optimal solution when the number $\sharp Arms$ of arms is large enough.

applications, we believe DeCo could be useful for decision setting where the horizon is short and the stack is high.

It should be noted that, as of now, the approach main limitations is that the horizon T is supposed to be known in advance and to be reasonably small (in the experiments, $T \leq 500$), whereas many MABs algorithms do not require T as an input. In addition, the usage of dynamic programming might make DeCo too burdensome for some applications. Also, since the DeCo algorithm requires a Bayesian prior, the question of the impact of a wrong prior on the performance is left open. On the other hand, DeCo can deal with time varying reward functions and can even include a final reward. In particular, DeCo can be applied to nonstationary settings where FH-Gittins cannot.

On the one hand, this work demonstrates that decomposition methods could be envisioned for addressing the exploration-exploitation tradeoff. On the other hand, bandit problems provide an interesting playground to extend the understanding of those methods, as demonstrated by Sect. 4. The characterization of the optimality gap presented in this paper, while focused on a specific problem, is a step toward better understanding decomposition methods. While it does not fully explain why DeCo works so well, it is a step into that direction as it provides a visual interpretation of the gap between the upper bound provided by the relaxed solution and the performance of DeCo.

References

[1] D. Adelman and A. J. Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727, 2008.

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

[3] D. Baudry, Y. Russac, and A. Filiot. Information_directed_sampling. https://github.com/DBaudry/, 2019.

[4] R. Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[6] D. B. Brown and J. E. Smith. Index policies and performance bounds for dynamic selection problems. *Management Science*, 66(7):3029–3050, 2020.

[7] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[8] P. Carpentier, J.-P. Chancelier, M. De Lara, and F. Pacaud. Mixed spatial and temporal decompositions for large-scale multistage stochastic optimization problems. *Journal of Optimization Theory and Applications*, 186(3):985–1005, 2020.

[9] P. Carpentier, J.-P. Chancelier, V. Leclère, and F. Pacaud. Stochastic decomposition applied to large-scale hydro valleys management. *European Journal of Operational Research*, 270(3):1086–1098, 2018.

[10] J. Chakravorty and A. 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):455, 2014.

[11] J.-P. Chancelier. Website: `http://cermics.enpc.fr/~jpc/nsp-tiddly`. Nsp, a numerical computing environment, 2021.

[12] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[13] V. F. Farias and E. Gutin. Optimistic Gittins indices. In *Proceedings of the 30th International Conference on Neural Information Processing Systems,(3161-3169)*, 2016.

[14] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[15] J. C. Gilbert and X. Jonsson. LIBOPT – An environment for testing solvers on heterogeneous collections of problems, 2007.

[16] J. Ginebra and M. K. Clayton. Small-sample performance of Bernoulli two-armed bandit Bayesian strategies. *Journal of statistical planning and inference*, 79(1):107–122, 1999.

[17] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[18] J. Gittins and Y.-G. Wang. The Learning Component of Dynamic Allocation Indices. *The Annals of Statistics*, 20(3):1625 – 1636, 1992.

[19] T. Hawkins. *A Lagrangian Decomposition Approach to Weakly Coupled Dynamic Optimization Problems and its Applications*. PhD thesis, Massachusetts Institute of Technology. Operations Research Center, 2003.

[20] P. Jacko. Binarybandit: An efficient Julia package for optimization and evaluation of the finite-horizon bandit problem with binary responses, 2019.

[21] P. Jacko. The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths, 2019.

[22] E. Kaufmann. On Bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842 – 865, 2018.

[23] T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

[24] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[25] T. Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. In *Conference on Learning Theory*, pages 1214–1245. PMLR, 2016.

[26] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[27] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[28] J. Nino-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

[29] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[30] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[31] W. 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.

[32] R. Weber. On the Gittins Index for Multiarmed Bandits. *The Annals of Applied Probability*, 2(4):1024 – 1033, 1992.

[33] J. Weed, V. Perchet, and P. Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.

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

[35] X. Zhang and P. I. Frazier. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv*, 2022.