Article

Multi-Gear Bandits, Partial Conservation Laws, and Indexability †

José Niño-Mora

Department of Statistics, Carlos III University of Madrid, 28903 Getafe, Spain; jose.nino@uc3m.es

† This paper was presented by the author at the 32nd European Conference on Operational Research (EURO 2022), Espoo, Finland, 3–6 July 2022. An earlier preliminary draft was presented by the author at the 3rd International Conference on Performance Evaluation Methodologies and Tools (ValueTools 2008), Athens, Greece, 20–24 October 2008, where an extended abstract was published in its proceedings.

Abstract: This paper considers what we propose to call *multi-gear bandits*, which are Markov decision processes modeling a generic dynamic and stochastic *project* fueled by a single resource and which admit multiple actions representing gears of operation naturally ordered by their increasing resource consumption. The optimal operation of a multi-gear bandit aims to strike a balance between project performance costs or rewards and resource usage costs, which depend on the resource price. A computationally convenient and intuitive optimal solution is available when such a model is *indexable*, meaning that its optimal policies are characterized by a *dynamic allocation index* (DAI), a function of state–action pairs representing critical resource prices. Motivated by the lack of general indexability conditions and efficient index-computing schemes, and focusing on the infinite-horizon finite-state and -action discounted case, we present a verification theorem ensuring that, if a model satisfies two proposed *PCL-indexability conditions* with respect to a postulated family of structured policies, then it is indexable and such policies are optimal, with its DAI being given by a *marginal productivity index* computed by a downshift adaptive-greedy algorithm in AN steps, with $A + 1$ actions and N states. The DAI is further used as the basis of a new index policy for the *multi-armed multi-gear bandit problem*.

Keywords: Markov decision process; multi-gear bandits; index policies; indexability; index algorithm

MSC: 90C40; 90C39; 90B36

1. Introduction

There is a substantial literature analyzing the *indexability* of infinite-horizon discrete-time binary-action Markov decision processes (MDP), i.e., their optimal solution by *index policies*, starting with the seminal work of Bellman in [1] on a Bernoulli bandit model. Such MDPs can be interpreted as models of a generic dynamic and stochastic *project*, which can be operated in a passive or an active mode. We propose to refer to such operating modes as *gears*, to reflect their natural ordering by increasing activity level.

The model in [1] had the property that, while the project is passive, its state does not change, which corresponds to a *classic bandit* setting. It was later shown that, given a finite collection of classic bandits, one of which must be active at each time, the policy that maximizes the expected total discounted reward in such a *multi-armed bandit problem* has a remarkably simple structure which overcomes the *curse of dimensionality* of a standard *dynamic programming* approach: it suffices to evaluate what Gittins and Jones [2] called the *dynamic allocation index* (DAI), later known as the *Gittins index*, of each project, which is a function of its state, and then activate at each time a project with largest index. See, e.g., the seminal work of Gittins and Jones [2] and Gittins [3], the monograph by Gittins [4], and alternative proofs by Whittle [5], Weber [6], and Bertsimas and Niño-Mora [7].

The assumption in classic bandit models that passive projects do not change state was removed by Whittle in [8], introducing *restless bandits*. That paper further introduced an

index for restless bandits (the *Whittle index*), which characterizes the optimal operation of a single restless project and provides a suboptimal heuristic policy for scheduling multiple such projects, in the so-called *multi-armed restless bandit problem*. The latter has huge modeling power but is computationally intractable, and Whittle's index policy has proven effective in an ever-increasing variety of models for multifarious applications. Thus, e.g., to name a few, scheduling multi-class make-to-stock queues [9], scheduling multi-class queues with finite buffers [10], admission control and routing to parallel queues with reneging [11], obsolescence mitigation strategies [12], sensor scheduling and dynamic channel selection [13–16], group maintenance [17], multi-target tracking with Kalman filter dynamics [18,19], scheduling multi-armed bandits with switching costs [20] or switching delays [21], the dynamic prioritization of medical treatments or interventions [22,23], and resource allocation with varying requests and with resources shared by multiple requests [24].

Yet, while the Gittins index is well defined for any classic bandit, the Whittle index only exists for some restless bandits, called *indexable*. Whittle pointed out in [8] the need of finding sufficient conditions for *indexability* (the existence of the index).

While researchers have deployed a wide variety of ingenious ad hoc techniques for proving the indexability of particular restless bandit models and computing their Whittle indices, the author has developed over the last two decades in a series of papers a systematic approach to accomplish such goals. Thus, Ref. [25] introduced a framework for establishing *both* the indexability of a general finite-state restless bandit *and* the optimality of a postulated family of structured policies, based on the concept of *partial conservation laws* (PCLs), also introduced there. If project performance metrics satisfy so-called *PCL-indexability conditions*, then the project's Whittle index can be efficiently computed in N steps, where N is the number of states, by an *adaptive-greedy index algorithm*. This algorithm is an extension of the classic index-computing algorithm of Klimov [26] for computing the indices that characterize the optimal policy for scheduling a multi-class queue with feedback. Note that Klimov's algorithm was adapted in [7] for computing the Gittins index, based on a framework of generalized conservation laws.

PCLs extend classical conservation laws in stochastic scheduling. See, e.g., the conservation laws in Coffman and Mitrani [27], the *strong conservation laws* in Shanthikumar and Yao [28], and the *generalized conservation laws* in the work of Bertsimas and Niño-Mora [7].

The author further developed the PCL framework for analyzing the indexability of finite-state restless bandits in [29], which introduced projects fueled by a generic resource with a general resource consumption function. The framework was extended to the countably infinite state space case in [30] and to the *bias optimality criterion* in [10]. Such early work is surveyed in the discussion paper [31]. The framework was then extended to projects with a continuous real state in [32], motivated by sensor scheduling applications. As for the adaptive-greedy algorithm for the Whittle index, an efficient computational implementation was presented in [33].

The extension of the concept of indexability from two-gear bandits to multi-gear bandits was first outlined by Weber [34] in the setting of an illustrative example given by a three-action queueing admission control model. Ref. [34] further outlined an index-computing algorithm for such a three-action model extending the aforementioned adaptive-greedy algorithm for the Whittle index. Such an insightful outline, was, however, not theoretically supported.

The author formalized and extended Weber's [34] outline to introduce in [35] a general multi-armed multi-mode bandit problem with finite-state projects, motivated by a model of optimal dynamic power allocation to multiple users sharing a wireless downlink communication channel subject to a peak energy constraint, proposing an index policy when individual projects are indexable. Ref. [35] further outlined an extension of the PCL framework along with an index algorithm, yet without proofs or analyses.

See also the recent work of Zayas-Cabán et al. [36] on a finite-horizon multi-armed multi-gear bandit problem and that of Killian et al. [37] on multi-action bandits, yet without a focus on indexability.

A related strand of research is the work on the optimality of structured policies in MDP models, mostly focusing on the monotonicity of optimal actions on the state. See, e.g., Serfozo [38] and the book chapter by Heyman and Sobel ([39] Ch. 8). In some models with a one-dimensional state, most notably those arising in queueing theory, researchers have established the optimality of policies given by multiple thresholds. Thus, e.g., Crabill [40] shows that such policies are optimal for selecting the speed of a single server catering to a queue. Such a model is an example of what we call here a *multi-gear bandit*. Similar results are obtained, e.g., in Sabeti [41], Ata and Shneorson [42], and Mayorga et al. [43]. The methods proposed herein also serve to establish the optimality of postulated families of structured policies, different from prevailing approaches, typically based on *submodularity*.

Another related line of work is the computational complexity of solving discounted finite-state and -action MDPs with general-purpose algorithms, most notably the classical methods of *value iteration*, *policy iteration*, and *linear optimization*. Such methods are not *strongly polynomial* in that the number of iterations required to compute an optimal policy depends not only on the number of states and actions but also on other factors, most notably the discount factor. Thus, e.g., Ye [44] showed that the number of iterations required by policy iteration is bounded by $O((N^2A/(1-\beta))\log(N/(1-\beta)))$, which shows that policy iteration is strongly polynomial but only if the discount factor is fixed and hence not part of the input. Such a bound has been improved by Scherrer [45] to $O((NA/(1-\beta))\log(1/(1-\beta)))$. Otherwise, policy iteration is known to have exponential worst-case complexity. See Hollanders et al. [46] and references therein. In contrast, as we will see, the algorithm presented herein solves a *multi-gear bandit* model with N states and $A+1$ actions in precisely AN steps and is hence a strongly polynomial algorithm.

In contrast with the aforementioned outlines in the earlier work [34,35], this paper presents a theoretically supported extension of the PCL-based sufficient indexability conditions from two-gear bandits to multi-gear bandits, along with an intuitive and efficient algorithm for computing the model's index.

The main contribution is a *verification theorem* (Theorem 1) which ensures that, if the performance metrics of a multi-gear bandit model under a postulated family of structured policies satisfy two *PCL-indexability* conditions, then *both* the model is indexable *and* such policies are optimal, with the model's index being computed by a *downshift adaptive-greedy algorithm* in AN steps as pointed out above.

The remainder of the paper is organized as follows. Section 2 describes the multi-gear bandit model and formulates the main result, the verification theorem for indexability. Sections 3–6 lay out the groundwork needed to prove the verification theorem. Thus, Section 3 discusses the linear optimization reformulation of the relevant MDP model. Section 4 presents the required relations between project performance metrics. Section 5 analyzes the output of the proposed index-computing algorithm. Section 6 presents the framework of partial conservation laws in the present setting. Then, Section 7 draws on the above to present our proof of the verification theorem. Section 8 applies the indexability property to provide a performance bound and a novel index policy for the multi-armed multi-gear bandit problem. Section 9 outlines some extensions, in particular to the long-run average cost criterion (Section 9.1), to models with uncontrollable states (Section 9.2) and to models with a countably infinite state space (Section 9.3). Finally, Section 10 concludes with a discussion of the results.

2. Preliminaries and Formulation of the Main Result

2.1. Multi-Gear Bandits

We next describe a general MDP model for the optimal operation of a multi-gear dynamic and stochastic project, which we call the *multi-gear bandit problem*. Consider a general discrete-time infinite-horizon discounted MDP model of a controlled dynamic

and stochastic *project* that consumes a single resource. At the start of each time period $t = 0, 1, \ldots$, the controller observes the current project *state* $s(t)$, which moves through the finite state space $\mathcal{N} \triangleq \{1, \ldots, N\}$, and then selects an *action* $a(t)$ from the finite action space $\mathcal{A} = \{0, 1, \ldots, A\}$. The choice of action at each time t is based on a possibly randomized function of the system's *history* $\mathcal{H}(t) \triangleq \{s(t)\} \cup \{(s(t'), a(t')) \colon t' = 0, \ldots, t-1\}$, consisting of the current state $s(t)$ and previous states visited and actions taken, if any. This corresponds to adopting a *control policy* (*policy* for short) π from the class Π of *history-dependent randomized policies* (see ([47] Sec. 2.1.5)). We will call such policies *admissible*.

We will refer to action 0 as the *passive* action, as it models the project's evolution in the absence of control, and to $1, \ldots, A$ as the *active* actions. Such actions model distinct *gears* for operating the project which are naturally ordered by their increasing resource consumption. Henceforth, we will use the terms *action* and *gear* interchangeably.

When the project occupies state $s(t) = i$ at the start of a period and action $a(t) = a$ is selected, it incurs a holding cost h_i^a and consumes a quantity q_i^a of the resource in the period, time-discounted with factor $0 < \beta < 1$. Then, the project state moves in a Markovian fashion from $s(t) = i$ to $s(t+1) = j$ with probability p_{ij}^a.

Consistently with the interpretation of actions a as operating gears ordered by increasing activity levels, we shall assume that higher gears entail larger resource consumptions, so the resource consumption q_i^a is monotone increasing in the gear a for each state i:

$$0 \leqslant q_i^0 < q_i^1 < \cdots < q_i^A, \quad i \in \mathcal{N}. \tag{1}$$

Intuitively, to compensate for their larger resource consumptions, higher gears should be more beneficial in some sense than lower gears, e.g., they might tend to drive the project towards less costly states or yield lower holding costs.

We further introduce a scalar parameter $\lambda \in \mathbb{R}$ modeling the resource unit price. Note that λ could take negative values, in which case it would represent a subsidy for using the resource. We shall consider the project's *λ-price problem*, which is to find an admissible project operating policy $\pi^*(\lambda)$ minimizing the expected total discounted holding and resource usage cost for *any* initial state. Writing as $\mathbb{E}_i^\pi[\cdot]$ the expectation under policy π starting from state i, we denote by

$$V_i(\lambda, \pi) \triangleq \mathbb{E}_i^\pi \left[\sum_{t=0}^\infty \left(h_{s(t)}^{a(t)} + \lambda q_{s(t)}^{a(t)} \right) \beta^t \right] \tag{2}$$

the corresponding expected total discounted cost incurred by the project when resource usage is charged at price λ. The resulting *optimal (project) cost function* is

$$V_i^*(\lambda) \triangleq \inf \{ V_i(\lambda, \pi) \colon \pi \in \Pi \}, \quad i \in \mathcal{N}.$$

We can thus formulate the project's λ-price problem as

$$(P_\lambda) \quad \text{find } \pi^*(\lambda) \in \Pi \colon V_i(\lambda, \pi^*(\lambda)) = V_i^*(\lambda), \quad i \in \mathcal{N}. \tag{3}$$

We shall refer to a policy $\pi^*(\lambda)$ solving the λ-price problem (P_λ) as a *λ-optimal policy*.

Denoting by Π^{SD} the class of *stationary deterministic policies* (see [47] (Sec. 2.1.5)), which base action choice on the current state only, standard results in MDP theory (see [47] (Theorem 6.2.10.a)) ensure the existence of a λ-optimal policy $\pi^*(\lambda) \in \Pi^{\text{SD}}$. Both the optimal cost function $V_i^*(\lambda)$ and the optimal stationary deterministic policies for the λ-price problem (P_λ) are determined by Bellman's *discounted-cost optimality equations*

$$V_i^*(\lambda) = \min_{a \in \mathcal{A}} \left(h_i^a + \lambda q_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a V_j^*(\lambda) \right), \quad i \in \mathcal{N}. \tag{4}$$

It is well known that the optimal cost function $V_i^*(\lambda)$ is the unique solution to such equations and that a stationary deterministic policy is optimal iff (i.e., if and only if) it selects an action attaining the minimum in the right-hand side of (4) for each state i. We shall also call such actions λ-optimal. For fixed λ, the Bellman equations can be solved numerically by classical methods, such as *value iteration*, *policy iteration*, and *linear optimization*. See, e.g., ([47] Sec. 6).

2.2. Indexability

Instead of solving the λ-price problem for specific values of the parameter λ, we shall pursue an alternative approach aiming at a complete understanding of optimal policies over the entire parameter space, by fully characterizing the optimal actions for the parametric collection $\mathcal{P} \triangleq \{P_\lambda : \lambda \in \mathbb{R}\}$ of *all* λ-price problems. Such a characterization will be given in terms of *critical parameter values* $\lambda_i^{*,a}$, as defined next.

Note that, below and throughout the paper, we use the standard abbreviation *iff* for *if and only if*.

Definition 1 (Indexability and DAI). *We call the above multi-gear bandit model indexable if there exist critical resource prices $\lambda_i^{*,a}$ for every state i and active action (gear) $a \geqslant 1$ satisfying $\lambda_i^{*,A} \leqslant \cdots \leqslant \lambda_i^{*,1}$, such that, for any such state and resource price $\lambda \in \mathbb{R}$: (i) action 0 is λ-optimal in state i iff $\lambda \geqslant \lambda_i^{*,1}$; (ii) action $1 \leqslant a \leqslant A - 1$ is λ-optimal in state i iff $\lambda_i^{*,a+1} \leqslant \lambda \leqslant \lambda_i^{*,a}$; and (iii) action A is λ-optimal in state i iff $\lambda \leqslant \lambda_i^{*,A}$. We call $\lambda_i^{*,a}$ the model's dynamic allocation index (DAI), viewed as a function of (i, a).*

Remark 1.

(i) *The definition of indexability for multi-action bandits was first outlined by Weber [34] in the setting of a three-action project model and was first formalized by the author [35] in the general setting considered herein. The latter paper further introduced the multi-armed multi-mode bandit problem and proposed to use the above DAI as the basis for a heuristic index policy for it. The concept of indexability for two-gear (active/passive) bandits has its roots in the work of Bellman [1], where he characterized the optimal policies for operating a Bernoulli bandit in terms of critical parameter values. Gittins and Jones [2] showed that such critical values (later known as Gittins indices) provide a tractable optimal policy for the classic multi-armed bandit problem, involving the optimal sequential activation of a collection of two-gear bandits that do not change state when passive. The idea of indexability was extended by Whittle [8] to two-gear bandits that can change state when passive, called restless bandits. He also proposed to use the corresponding Whittle index policy as a heuristic for the intractable multi-armed restless bandit problem, when the individual bandits (projects) are indexable, which need not be the case as there are nonindexable bandits.*

(ii) *Writing as $V_i(\lambda, \langle a, * \rangle) \triangleq h_i^a + \lambda q_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a V_j^*(\lambda)$ the optimal cost function with initial action a, indexability means that there exist critical prices $\lambda_i^{*,a}$ as in Definition 1 such that, for each state i,*

$$V_i(\lambda, \langle 0, * \rangle) \leqslant V_i(\lambda, \langle a, * \rangle) \text{ for } a \geqslant 1 \iff \lambda \geqslant \lambda_i^{*,1}$$

$$\text{for } 0 < a < A : V_i(\lambda, \langle a, * \rangle) \leqslant V_i(\lambda, \langle a', * \rangle) \text{ for } a' \neq a \iff \lambda_i^{*,a+1} \leqslant \lambda \leqslant \lambda_i^{*,a} \quad (5)$$

$$V_i(\lambda, \langle A, * \rangle) \leqslant V_i(\lambda, \langle a, * \rangle) \text{ for } a < A \iff \lambda \leqslant \lambda_i^{*,A}.$$

(iii) *In intuitive terms, when an indexable project model occupies state i, it is λ-optimal to select the lowest (passive) gear 0 iff the resource is expensive enough ($\lambda \geqslant \lambda_i^{*,1}$); it is λ-optimal to select the highest gear A iff the resource is cheap enough ($\lambda \leqslant \lambda_i^{*,a}$); and it is λ-optimal to select the intermediate gear $0 < a < A$ iff the resource price lies between the critical prices $\lambda_i^{*,a+1}$ and $\lambda_i^{*,a}$.*

(iv)　*In an indexable model, $\lambda_i^{*,a}$ is the unique critical resource price λ for which gears $a - 1$ and a are both λ-optimal in state i; hence, it is the unique solution to the equation*

$$V_i(\lambda, \langle a - 1, * \rangle) = V_i(\lambda, \langle a, * \rangle). \tag{6}$$

　　Yet, note that for a nonindexable model, Equation (6) need not have a unique solution.

(v)　*Let $\mathcal{A}_i^*(\lambda)$ be the set of λ-optimal actions in state i. If the model is indexable then, for each active action $a \geqslant 1$, $\mathcal{A}_i^*(\lambda) \cap \{a, \ldots, A\} \neq \varnothing$ (i.e., there is an optimal action greater than or equal to a) iff $\lambda_i^{*,a} \geqslant \lambda$ and $\mathcal{A}_i^*(\lambda) \cap \{0, \ldots, a - 1\} \neq \varnothing$ (i.e., there is an optimal action less than a) iff $\lambda_i^{*,a} \leqslant \lambda$.*

We shall address the following research goals: (1) identify sufficient conditions ensuring that the above multi-gear bandit model is indexable and (2) for models satisfying such conditions, provide an efficient means of computing the DAI.

2.3. Project Performance Metrics and Their Characterization

To formulate the main result and facilitate the required analyses, we consider certain project *performance metrics*. We measure the holding cost incurred by the project under a policy $\pi \in \Pi$ starting from the initial-state distribution $s(0) \sim p = (p_i)_{i \in \mathcal{N}}$, so $\mathbb{P}\{s(0) = i\} = p_i$ for $i \in \mathcal{N}$, by the *(holding) cost metric*

$$F_p(\pi) = \mathbb{E}_p^\pi \left[\sum_{t=0}^\infty h_{s(t)}^{a(t)} \beta^t \right],$$

where $\mathbb{E}_p^\pi[\cdot]$ denotes expectation under policy π starting from $s(0) \sim p$. Similarly, we measure the corresponding resource usage by the *resource (usage) metric*

$$G_p(\pi) = \mathbb{E}_p^\pi \left[\sum_{t=0}^\infty q_{s(t)}^{a(t)} \beta^t \right].$$

When $s(0) = i$, we write $F_i(\pi)$ and $G_i(\pi)$. Note that $F_p(\pi) = \sum_i p_i F_i(\pi)$ and $G_p(\pi) = \sum_i p_i G_i(\pi)$. As for the total cost metric $V_i(\lambda, \pi)$ in (2), we can thus express it as

$$V_i(\lambda, \pi) = F_i(\pi) + \lambda G_i(\pi).$$

We shall similarly write $V_p(\lambda, \pi)$ when $s(0) \sim p$.

We next address the characterization of such metrics for stationary deterministic policies. We will represent any such policy by the partition $S = (S_a)_{a \in \mathcal{A}} = (S_0, \ldots, S_A)$ it naturally induces on the state space $\mathcal{N}$, where S_a is the subset of states where the policy selects gear a. We shall refer to it as the *S-policy* or *policy S*.

The performance metrics $F_i(S)$ and $G_i(S)$ for the S-policy are thus characterized as the unique solutions to the linear equation systems

$$F_i(S) = h_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a F_j(S), \quad i \in S_a, a \in \mathcal{A},$$

and

$$G_i(S) = q_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a G_j(S), \quad i \in S_a, a \in \mathcal{A}.$$

In the sequel we will find it convenient to use vector notation, denoting vectors and matrices in boldface and writing, e.g., $\mathbf{F}(S) = (F_i(S))_{i \in \mathcal{N}}$ and $\mathbf{F}_B(S) = (F_i(S))_{i \in B}$ for $B \subset \mathcal{N}$, and similarly for $\mathbf{G}(S)$, $\mathbf{h}^a$ and $\mathbf{q}^a$. We will also write $\mathbf{P}^a = (p_{ij}^a)_{i,j \in \mathcal{N}}$, $\mathbf{P}_{BB'}^a = (p_{ij}^a)_{i \in B, j \in B'}$ and $\mathbf{P}_{B\cdot}^a = (p_{ij}^a)_{i \in B, j \in \mathcal{N}}$ for $B, B' \subset \mathcal{N}$.

The above equations characterizing $\mathbf{F}(S)$ and $\mathbf{G}(S)$ are thus formulated as

$$\mathbf{F}_{S_a}(S) = \mathbf{h}_{S_a}^a + \beta \mathbf{P}_{S_a\cdot}^a \mathbf{F}(S), \quad a \in \mathcal{A}. \tag{7}$$

and

$$\mathbf{G}_{S_a}(S) = \boldsymbol{q}_{S_a}^a + \beta \mathbf{P}_{S_a}^a \cdot \mathbf{G}(S), \quad a \in \mathcal{A}. \tag{8}$$

We shall further consider corresponding marginal metrics. Denote by $\langle a, S \rangle$ the policy that selects gear a at time $t = 0$ and then adopts the S-policy thereafter. Note that

$$F_i(\langle a, S \rangle) = h_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a F_j(S)$$

and

$$G_i(\langle a, S \rangle) = q_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^a G_j(S).$$

For given actions $a \neq a'$, we define the *marginal (holding) cost metric*

$$f_i^{a,a'}(S) \triangleq F_i(\langle a, S \rangle) - F_i(\langle a', S \rangle) = h_i^a - h_i^{a'} + \beta \sum_{j \in \mathcal{N}} p_{ij}^a F_j(S) - \beta \sum_{j \in \mathcal{N}} p_{ij}^{a'} F_j(S), \tag{9}$$

which measures the *decrement* in the holding cost metric that results from the shifting of the initial gear from a to a' starting from state i, provided that the S-policy is followed thereafter.

We also define the *marginal resource (usage) metric*

$$g_i^{a,a'}(S) \triangleq G_i(\langle a', S \rangle) - G_i(\langle a, S \rangle) = q_i^{a'} - q_i^a + \beta \sum_{j \in \mathcal{N}} p_{ij}^{a'} G_j(S) - \beta \sum_{j \in \mathcal{N}} p_{ij}^a G_j(S), \tag{10}$$

which measures the corresponding *increment* in the resource metric.

In vector notation, we can write the above identities as

$$\mathbf{f}^{a,a'}(S) \triangleq \mathbf{F}(\langle a, S \rangle) - \mathbf{F}(\langle a', S \rangle) = \mathbf{h}^a - \mathbf{h}^{a'} + \beta(\mathbf{P}^a - \mathbf{P}^{a'})\mathbf{F}(S) \tag{11}$$

and

$$\mathbf{g}^{a,a'}(S) \triangleq \mathbf{G}(\langle a', S \rangle) - \mathbf{G}(\langle a, S \rangle) = \boldsymbol{q}^{a'} - \boldsymbol{q}^a + \beta(\mathbf{P}^{a'} - \mathbf{P}^a)\mathbf{G}(S). \tag{12}$$

If $g_i^{a,a'}(S) > 0$ for certain i, a, a', and S, we further define the *marginal productivity (MP) metric* as the ratio of the marginal cost metric to the marginal resource metric:

$$m_i^{a,a'}(S) \triangleq \frac{f_i^{a,a'}(S)}{g_i^{a,a'}(S)}. \tag{13}$$

We next present a preliminary result, on which we draw later on, establishing further relations between metrics $\mathbf{F}(S)$ and $\mathbf{f}^{a,a'}(S)$ and between $\mathbf{G}(S)$ and $\mathbf{g}^{a,a'}(S)$. Note that, below, $\mathbf{0}_{S_a}$ denotes a vector of zeros with components indexed by S_a.

Lemma 1. *For any stationary deterministic policy S and action a,*

(a) $\quad (\mathbf{I} - \beta \mathbf{P}^a)\mathbf{F}(S) - \mathbf{h}^a = \begin{bmatrix} (\mathbf{f}_{S_{a'}}^{a',a}(S))_{a' \in \mathcal{A} - \{a\}} \\ \mathbf{0}_{S_a} \end{bmatrix};$

(b) $\quad \boldsymbol{q}^a - (\mathbf{I} - \beta \mathbf{P}^a)\mathbf{G}(S) = \begin{bmatrix} (\mathbf{g}_{S_{a'}}^{a',a}(S))_{a' \in \mathcal{A} - \{a\}} \\ \mathbf{0}_{S_a} \end{bmatrix}.$

Proof. (a) For $a' \neq a$, using in turn (7) and (11) we obtain

$$\begin{aligned}
\mathbf{f}_{S_{a'}}^{a',a}(S) &= \mathbf{h}_{S_{a'}}^{a'} - \mathbf{h}_{S_{a'}}^a + \beta(\mathbf{P}_{S_{a'}}^{a'} - \mathbf{P}_{S_{a'}}^a)\mathbf{F}(S) \\
&= (\mathbf{F}_{S_{a'}}(S) - \mathbf{h}_{S_{a'}}^{a'} - \beta\mathbf{P}_{S_{a'}}^{a'}\mathbf{F}(S)) + (\mathbf{h}_{S_{a'}}^{a'} - \mathbf{h}_{S_{a'}}^a + \beta(\mathbf{P}_{S_{a'}}^{a'} - \mathbf{P}_{S_{a'}}^a)\mathbf{F}(S)) \\
&= \mathbf{F}_{S_{a'}}(S) - \beta\mathbf{P}_{S_{a'}}^a\mathbf{F}(S) - \mathbf{h}_{S_{a'}}^a,
\end{aligned}$$

and

$$\mathbf{F}_{S_a}(S) - \beta \mathbf{P}^a_{S_a} \cdot \mathbf{F}(S) - \mathbf{h}^a_{S_a} = \mathbf{0}_{S_a}.$$

Part (b) follows similarly. □

2.4. Main Result: A Verification Theorem for Indexability

We next present our main result, giving sufficient conditions for indexability. The conditions correspond to the framework of *PCL-indexability*, which is extended here from the two-gear setting in [25,29,30,32] to the multi-gear setting.

We will use the following notation. Given a stationary deterministic policy $S = (S_0, \ldots, S_A)$, actions $a \neq a'$, and a state $j \in S_a$, the policy denoted by $\widehat{S} = \mathcal{T}^{a,a'}_j S$ is defined by $\widehat{S}_a = S_a - \{j\}$, $\widehat{S}_{a'} = S_{a'} \cup \{j\}$, and $\widehat{S}_{a''} = S_{a''}$ for $a \neq a'' \neq a'$. Thus, $\mathcal{T}^{a,a'}_j S$ is obtained from S by shifting the gear selected in state j from a to a'.

The verification theorem below refers to indexability relative to a structured family $\mathcal{F}$ of stationary deterministic policies, which one needs to postulate a priori, based on insight on the particular model at hand. We shall thus refer to the family of *$\mathcal{F}$-policies* $S \in \mathcal{F}$, and to *$\mathcal{F}$-indexability*, as defined below.

Definition 2 ($\mathcal{F}$-indexability). *We call the model $\mathcal{F}$-indexable if (i) it is indexable and (ii) $\mathcal{F}$-policies are optimal for the λ-price problem (P_λ) in (3), for any $\lambda \in \mathbb{R}$.*

Note that Definition 2(ii) refers to the *optimality of $\mathcal{F}$-policies* for all λ-price problems (P_λ). By this we mean that, for any $\lambda \in \mathbb{R}$, there exists a λ-optimal policy $S^*(\lambda) \in \mathcal{F}$.

We require $\mathcal{F}$ to satisfy the following *connectedness assumption*, which is motivated by algorithmic considerations. The assumption ensures that it is possible to go from policy $(\varnothing, \ldots, \varnothing, \mathcal{N})$ to $(\mathcal{N}, \varnothing, \ldots, \varnothing)$, both of which must be in $\mathcal{F}$, through a sequence of policies in $\mathcal{F}$ where each policy in the sequence is obtained from the previous one by *downshifting* the gear selected in a single state to the next lower gear. Conversely, it is possible to go from $(\mathcal{N}, \varnothing, \ldots, \varnothing)$ to $(\varnothing, \ldots, \varnothing, \mathcal{N})$ through a sequence of policies in $\mathcal{F}$ where each policy in the sequence is obtained from the previous one by *upshifting* the gear selected in a single state to the next higher gear.

Assumption 1. *The family of policies $\mathcal{F}$ satisfies the following conditions:*

(i) $(\varnothing, \ldots, \varnothing, \mathcal{N}) \in \mathcal{F}$ *and* $(\mathcal{N}, \varnothing, \ldots, \varnothing) \in \mathcal{F}$;

(ii) *For each* $S \in \mathcal{F} - \{(\mathcal{N}, \varnothing, \ldots, \varnothing)\}$ *there exist* $a \geqslant 1$ *and* $j \in S_a$ *such that* $\mathcal{T}^{a,a-1}_j S \in \mathcal{F}$;

(iii) *For each* $S \in \mathcal{F} - \{(\varnothing, \ldots, \varnothing, \mathcal{N})\}$ *there exist* $a < A$ *and* $j \in S_a$ *such that* $\mathcal{T}^{a,a+1}_j S \in \mathcal{F}$.

Note that the above concepts of downshifting and upshifting gears naturally induce a *partial ordering* $\preceq$ on the class of all stationary deterministic policies S and in particular on $\mathcal{F}$. Thus, given S and S', we write $S \preceq S'$ if, at every state, S does not select a higher gear than S'. If, further, $S \neq S'$, we write $S \prec S'$. Assumption 1 shows that the *poset* (partially ordered set) $(\mathcal{F}, \preceq)$ contains the least element $(\mathcal{N}, \varnothing, \ldots, \varnothing)$ and the largest element $(\varnothing, \ldots, \varnothing, \mathcal{N})$.

The verification theorem refers to the *downshift adaptive-greedy index algorithm* $\mathrm{DS}(\mathcal{F})$ shown in Algorithm 1. This takes as input the model parameters and, in $K \triangleq AN$ steps, produces as output a sequence of distinct state–action pairs (j_k, a_k) spanning $\mathcal{N} \times (\mathcal{A} - \{0\})$ along with corresponding sequences of $\mathcal{F}$-policies S^k and scalars $m^{*,a_k}_{j_k}$ for $k = 1, \ldots, K$. Actually, the sequence S^k is a *chain* of the poset $(\mathcal{F}, \preceq)$ ordered as

$$(\mathcal{N}, \varnothing, \ldots, \varnothing) = S^{K+1} \prec S^K \prec \cdots \prec S^1 = (\varnothing, \ldots, \varnothing, \mathcal{N}).$$

Algorithm 1: Downshift adaptive-greedy index algorithm DS($\mathcal{F}$).

Output: $\{(j_k, a_k), S^k, m_{j_k}^{*,a_k}\}_{k=1}^{K}$

Initialization: $S^1 := (\varnothing, \ldots, \varnothing, \mathcal{N})$; $a_1 := A$
pick $j_1 \in \arg\min_{j \in \mathcal{N}, \mathcal{T}_j^{a_1, a_1 - 1} S^1 \in \mathcal{F}} m_j^{a_1 - 1, a_1}(S^1)$
$m_{j_1}^{*, a_1} := m_{j_1}^{a_1 - 1, a_1}(S^1)$; $S^2 := \mathcal{T}_{j_1}^{a_1, a_1 - 1} S^1$
Loop:
for $k := 2$ **to** K **do**

    pick $(j_k, a_k) \in \arg\min_{(j,a): j \in S_a^k, \mathcal{T}_j^{a, a-1} S^k \in \mathcal{F}} m_j^{a-1, a}(S^k)$,

      with $m_j^{a-1, a}(S^k) := m_{j_{k-1}}^{*, a_{k-1}} + \dfrac{g_j^{a-1, a}(S^{k-1})}{g_j^{a-1, a}(S^k)}(m_j^{a-1, a}(S^{k-1}) - m_{j_{k-1}}^{*, a_{k-1}})$

    $m_{j_k}^{*, a_k} := m_{j_k}^{a_k - 1, a_k}(S^k)$; $S^{k+1} := \mathcal{T}_{j_k}^{a_k, a_k - 1} S^k$
end { for }

Remark 2.

(i) *Algorithm* DS($\mathcal{F}$) *extends to multi-gear bandits the adaptive-greedy algorithm for computing the Whittle index for restless (two-gear) bandits introduced by the author in [25] and further developed in [29]. In turn, this has its early roots in Klimov's adaptive-greedy index algorithm [26] for computing the indices that give the optimal policy for scheduling a multi-class queue with feedback. Note that Klimov's algorithm was first adapted in [7] to compute the Gittins index for classic bandits (i.e., two-gear bandits that do not change state when passive).*

(ii) *The term downshift refers to the way in which the algorithm generates the sequence S^k of $\mathcal{F}$-policies. It starts with policy S^1, which selects the highest gear A in every state. Then, at each step k of the algorithm, it selects a state j_k in which to downshift gears from a_k to $a_k - 1$, keeping the same gears in the other states, thus obtaining the next policy S^{k+1}. The selection of such a state j_k is performed in an adaptive-greedy fashion, by choosing a state in which such a downshifting change entails a minimal MP decrease, as measured by the MP metric $m_j^{a-1, a}(S^k)$, while also ensuring that the next policy S^{k+1} will be in $\mathcal{F}$. One can visualize the workings of the algorithm in terms of balls trickling down a grid in which states are positioned in columns and gears in rows, with gear A at the top. Initially, all balls are in the top row, as the algorithm starts with policy $S^1 \triangleq (\varnothing, \ldots, \varnothing, \mathcal{N})$, so gear A is chosen in every state. Then, at each step of the algorithm, one ball trickles down from its current row to that immediately below, which represents another $\mathcal{F}$-policy. The algorithm ends in K steps when all N balls have trickled down to the bottom row, which corresponds to policy $S^{K+1} \triangleq (\mathcal{N}, \varnothing, \ldots, \varnothing)$, so gear 0 is chosen in every state.*

(iii) *Note that, by construction, (1) each policy S^k in the sequence produced by the algorithm satisfies $S^k \in \mathcal{F}$, i.e., it is an $\mathcal{F}$-policy, for $k = 1, \ldots, K + 1$, and (2) the state–action pairs (j_k, a_k) produced by the algorithm are all distinct, spanning the K state–action pairs $(j, a) \in \mathcal{N} \times (\mathcal{A} - \{0\})$ corresponding to active gears $a \geqslant 1$.*

Definition 3 (PCL($\mathcal{F}$)-indexability and MP index)**.** *We call a multi-gear bandit model PCL-indexable with respect to $\mathcal{F}$-policies, or PCL($\mathcal{F}$)-indexable, if the following hold:*

(PCLI1) $g_j^{a-1, a}(S) > 0$ *for every policy $S \in \mathcal{F}$, active action $a \geqslant 1$, and state $j \in \mathcal{N}$;*

(PCLI2)   *Algorithm* DS($\mathcal{F}$) *computes the $m_{j_k}^{*, a_k}$ in nondecreasing order:*

$$m_{j_1}^{*, a_1} \leqslant m_{j_2}^{*, a_2} \leqslant \cdots \leqslant m_{j_K}^{*, a_K}.$$

In such a case, we call $m_j^{*, a}$ the project's MP index or MPI for short.

Remark 3.

(i) *Condition (PCLI1) means that the marginal resource metric corresponding to upshifting gears in any state, relative to any $\mathcal{F}$-policy, is positive. Note that it is equivalent to requiring that $g_j^{a',a}(S) > 0$ for $a' < a$, since $g_j^{a',a}(S) = g_j^{a',a'+1}(S) + \cdots + g_j^{a-1,a}(S)$. Such a condition and the fact that the S^k are in $\mathcal{F}$ ensures that the MP index $m_j^{*,a}$ computed by the algorithm is well defined.*

(ii) *The recursive formula used in the algorithm for computing the MP metrics as*

$$m_j^{a-1,a}(S^k) := m_{j_{k-1}}^{*,a_{k-1}} + \frac{g_j^{a-1,a}(S^{k-1})}{g_j^{a-1,a}(S^k)} \left(m_j^{a-1,a}(S^{k-1}) - m_{j_{k-1}}^{*,a_{k-1}} \right) \tag{14}$$

is justified by Lemma 8.

We next state the verification theorem.

Theorem 1. *If a multi-gear bandit model is PCL($\mathcal{F}$)-indexable, then it is $\mathcal{F}$-indexable with its DAI being given by its MPI, i.e., $\lambda_j^{*,a} = m_j^{*,a}$.*

The proof of Theorem 1 is presented in Section 7. It requires substantial preliminary groundwork, which is laid out in Sections 3–6.

3. Linear Optimization Reformulation of the λ-Price Problem

We start the required groundwork by reviewing the standard *linear optimization* (LO) formulation of a finite-state and -action MDP (see, e.g., ([47] Sec. 6.9)), since it applies to the λ-price problem (P_λ) in (3), as it is needed in subsequent analyses. It is well known that such an MDP can be reformulated as an LO problem on variables x_j^a for state–action pairs $(j, a) \in \mathcal{K} \triangleq \mathcal{N} \times \mathcal{A}$, which represent *discounted state–action occupancy measures*. Thus, variable x_j^a corresponds to the measure

$$x_{pj}^a(\pi) \triangleq \mathbb{E}_p^\pi \left[\sum_{t=0}^{\infty} \mathbf{1}_{\{(j,a)\}}(s(t), a(t)) \beta^t \right],$$

where $\mathbf{1}_B(\cdot)$ denotes the indicator function of a set B, so $x_{pj}^a(\pi)$ is the expected total discounted number of times that action a is selected in state j under policy π starting from $s(0) \sim p$. We write it as $x_{ij}^a(\pi)$ when $s(0) = i$.

Such a standard LO formulation is

$$(L_\lambda(p)): \qquad \text{minimize} \sum_{(j,a) \in \mathcal{K}} (h_j^a + \lambda q_j^a) x_j^a$$

$$\text{subject to: } x_j^a \geqslant 0 \tag{15}$$

$$\sum_{a \in \mathcal{A}} x_j^a - \beta \sum_{(i,a) \in \mathcal{K}} p_{ij}^a x_i^a = p_j, \ \ j \in \mathcal{N},$$

or, in vector notation, writing the probability mass function p as a vector $\mathbf{p}$,

$$(L_\lambda(p)): \qquad \text{minimize} \sum_{a \in \mathcal{A}} \mathbf{x}^a (\mathbf{h}^a + \lambda \mathbf{q}^a)$$

$$\text{subject to: } \mathbf{x}^a \geqslant \mathbf{0}, \ \ a \in \mathcal{A} \tag{16}$$

$$\sum_{a \in \mathcal{A}} \mathbf{x}^a (\mathbf{I} - \beta \mathbf{P}^a) = \mathbf{p}.$$

We next use the above LO formulation to show that the λ-price problem (P_λ) in (3) can be reformulated into an equivalent problem in which holding costs under action A are zero, a result on which we draw later on.

Thus, define

$$\widehat{\mathbf{h}}^a \triangleq \mathbf{h}^a - (\mathbf{I} - \beta \mathbf{P}^a)(\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A, \quad a \in \mathcal{A}, \tag{17}$$

and note that

$$\widehat{\mathbf{h}}^A = \mathbf{0}. \tag{18}$$

Denote by $\widehat{F}_i(\pi)$, $\widehat{f}_i^{a,a'}(S)$, and $\widehat{m}_i^{a,a'}(S)$ the cost, marginal cost, and marginal productivity metrics defined in Section 2.3, but for the model with modified holding costs $\widehat{h}_i^a$ in (17). The following result clarifies the relations between such metrics and those for the original holding costs h_i^a. Recall that we denote by $S^1 \triangleq (\varnothing, \ldots, \varnothing, \mathcal{N})$ the policy that selects the highest gear in every state.

Lemma 2. *For any admissible policy π, stationary deterministic policy S, and actions $a \neq a'$:*

(a) $\widehat{F}_p(\pi) = F_p(\pi) - F_p(S^1)$;

(b) $\widehat{f}_i^{a,a'}(S) = f_i^{a,a'}(S)$;

(c) $\widehat{m}_i^{a,a'}(S) = m_i^{a,a'}(S)$.

Proof. (a) The measures $\mathbf{x}_p^a(\pi) = (x_{pj}^a(\pi))_{j \in \mathcal{N}}$, viewed as row vectors, satisfy the constraints in (16):

$$\sum_{a=0}^{A} \mathbf{x}_p^a(\pi)(\mathbf{I} - \beta \mathbf{P}^a) = \mathbf{p}. \tag{19}$$

Hence, since the matrices $\mathbf{I} - \beta \mathbf{P}^a$ are invertible, we can write

$$\mathbf{x}_p^A(\pi) + \sum_{a=0}^{A-1} \mathbf{x}_p^a(\pi)(\mathbf{I} - \beta \mathbf{P}^a)(\mathbf{I} - \beta \mathbf{P}^A)^{-1} = \mathbf{p}(\mathbf{I} - \beta \mathbf{P}^A)^{-1}.$$

We thus obtain

$$
\begin{aligned}
F_p(\pi) = \sum_{a=0}^{A} \mathbf{x}_p^a(\pi)\mathbf{h}^a &= \sum_{a=0}^{A-1} \mathbf{x}_p^a(\pi)\mathbf{h}^a + \mathbf{x}_p^A(\pi)\mathbf{h}^A \\
&= \mathbf{p}(\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A + \sum_{a=0}^{A-1} \mathbf{x}_p^a(\pi)\big[\mathbf{h}^a - (\mathbf{I} - \beta \mathbf{P}^a)(\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A\big] \\
&= \mathbf{p}(\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A + \sum_{a=0}^{A-1} \mathbf{x}_p^a(\pi)\widehat{\mathbf{h}}^a \\
&= F_p(S^1) + \widehat{F}_p(\pi),
\end{aligned}
$$

where the last line is obtained from the previous one by using Equation (7), which yields

$$\mathbf{F}(S^1) = (\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A. \tag{20}$$

(b) We have, using (11) and (17), and part (a),

$$
\begin{aligned}
\widehat{\mathbf{f}}^{a,a'}(S) \triangleq \widehat{\mathbf{h}}^a - \widehat{\mathbf{h}}^{a'} + \beta(\mathbf{P}^a - \mathbf{P}^{a'})\widehat{\mathbf{F}}(S) \\
= \mathbf{h}^a - \mathbf{h}^{a'} + \beta(\mathbf{P}^a - \mathbf{P}^{a'})(\mathbf{I} - \beta \mathbf{P}^A)^{-1}\mathbf{h}^A + \beta(\mathbf{P}^a - \mathbf{P}^{a'})(\mathbf{F}(S) - \mathbf{F}(S^1)) \\
= \mathbf{h}^a - \mathbf{h}^{a'} + \beta(\mathbf{P}^a - \mathbf{P}^{a'})\mathbf{F}(S) \\
= \mathbf{f}^{a,a'}(S).
\end{aligned}
$$

(c) This part follows directly from part (b) and (13), since $\widehat{m}_i^{a,a'}(S) \triangleq \widehat{f}_i^{a,a'}(S)/g_i^{a,a'}(S) = f_i^{a,a'}(S)/g_i^{a,a'}(S) = m_i^{a,a'}(S)$. $\square$

Corollary 1. *For any state j,*

(a) $f_j^{a-1,a}(S^1) = \widehat{h}_j^{a-1} - \widehat{h}_j^a, \quad a \geqslant 1;$

(b) $m_j^{a,A}(S^1) = \widehat{h}_j^a / g_j^{a,A}(S^1), \quad a < A.$

Proof. (a) We have

$$\mathbf{f}^{a-1,a}(S^1) = \widehat{\mathbf{f}}^{a-1,a}(S^1) \triangleq \widehat{\mathbf{h}}^{a-1} - \widehat{\mathbf{h}}^a + \beta(\mathbf{P}^{a-1} - \mathbf{P}^a)\widehat{\mathbf{F}}(S^1) = \widehat{\mathbf{h}}^{a-1} - \widehat{\mathbf{h}}^a,$$

where we have used in turn Lemma 2(b), (11), (18) and (20).

(b) This part follows from (a), since $m_j^{a,A}(S^1) = f_j^{a,A}(S^1)/g_j^{a,A}(S^1) = \widehat{h}_j^a / g_j^{a,A}(S^1)$. □

Denote by $(\widehat{P}_\lambda)$ the modified λ-price problem where the $\mathbf{h}^a$ in problem (P_λ) are replaced by $\widehat{\mathbf{h}}^a$. Write as $\widehat{V}_p(\lambda, \pi)$ the project cost metric for the modified model.

Lemma 3. *Problems* (P_λ) *and* $(\widehat{P}_\lambda)$ *are equivalent, since* $\widehat{V}_p(\lambda, \pi) = V_p(\lambda, \pi) - F_p(S^1).$

Proof. We have, using Lemma 2(a),

$$\widehat{V}_p(\lambda, \pi) = \widehat{F}_p(\pi) + \lambda G_p(\pi) = V_p(\lambda, \pi) - F_p(S^1).$$

□

4. Relations between Performance Metrics

This section presents relations between performance metrics that we will need to prove the verification theorem. We start with a result giving decomposition identities that relate the metrics $F_p(\pi)$ and $G_p(\pi)$ for an admissible policy π to the metrics $F_p(S)$ and $G_p(S)$ under a particular stationary deterministic policy S, where p is the initial-state distribution. The resulting *decomposition identities* involve marginal cost metrics $f_j^{a,a'}(S)$ and marginal resource metrics $g_j^{a',a}(S)$.

Lemma 4 (Performance metrics decomposition). *For any admissible policy π and stationary deterministic policy S:*

(a) $F_p(S) + \sum_{a<a'} \sum_{j \in S_{a'}} f_j^{a,a'}(S) x_{pj}^a(\pi) = F_p(\pi) + \sum_{a'<a} \sum_{j \in S_{a'}} f_j^{a',a}(S) x_{pj}^a(\pi);$

(b) $G_p(\pi) + \sum_{a<a'} \sum_{j \in S_{a'}} g_j^{a,a'}(S) x_{pj}^a(\pi) = G_p(S) + \sum_{a'<a} \sum_{j \in S_{a'}} g_j^{a',a}(S) x_{pj}^a(\pi).$

Proof. (a) We can write, using in turn (19), Lemma 1(a), and $f_j^{a',a}(S) = -f_j^{a,a'}(S)$,

$$\begin{aligned}
0 &= \left[\sum_a \mathbf{x}_p^a(\pi)(\mathbf{I} - \beta\mathbf{P}^a) - \mathbf{p} \right] \mathbf{F}(S) = \sum_a \mathbf{x}_p^a(\pi)(\mathbf{I} - \beta\mathbf{P}^a)\mathbf{F}(S) - \mathbf{p}\mathbf{F}(S) \\
&= \sum_a \mathbf{x}_p^a(\pi)\left[(\mathbf{I} - \beta\mathbf{P}^a)\mathbf{F}(S) - \mathbf{h}^a \right] - \mathbf{p}\mathbf{F}(S) + \sum_a \mathbf{x}_p^a(\pi)\mathbf{h}^a \\
&= \sum_a \mathbf{x}_p^a(\pi)\left[(\mathbf{I} - \beta\mathbf{P}^a)\mathbf{F}(S) - \mathbf{h}^a \right] - F_p(S) + F_p(\pi) \\
&= \sum_{a' \neq a} \mathbf{x}_{pS_{a'}}^a(\pi)\mathbf{f}_{S_{a'}}^{a',a}(S) - F_p(S) + F_p(\pi) \\
&= \sum_{a'>a} \mathbf{x}_{pS_{a'}}^a(\pi)\mathbf{f}_{S_{a'}}^{a',a}(S) + \sum_{a'<a} \mathbf{x}_{pS_{a'}}^a(\pi)\mathbf{f}_{S_{a'}}^{a',a}(S) - F_p(S) + F_p(\pi) \\
&= \sum_{a'<a} \mathbf{x}_{pS_{a'}}^a(\pi)\mathbf{f}_{S_{a'}}^{a',a}(S) - \sum_{a'>a} \mathbf{x}_{pS_{a'}}^a(\pi)\mathbf{f}_{S_{a'}}^{a,a'}(S) - F_p(S) + F_p(\pi) \\
&= \sum_{a'<a} \sum_{j \in S_{a'}} x_{pj}^a(\pi)f_j^{a',a}(S) - \sum_{a'>a} \sum_{j \in S_{a'}} x_{pj}^a(\pi)f_j^{a,a'}(S) - F_p(S) + F_p(\pi).
\end{aligned}$$

Part (b) follows similarly as part (a). □

The following result draws on the above to relate metrics $F_p(S)$ and $F_p(\mathcal{T}_j^{a,a'} S)$ to $G_p(S)$ and $G_p(\mathcal{T}_j^{a,a'} S)$, respectively, which clarifies the interpretation of marginal metrics $f_j^{a,a'}(S)$ and $g_j^{a,a'}(S)$. Recall that $\mathcal{T}_j^{a,a'} S$ is the modification of policy S that results by shifting gear in state j from a to a'.

Lemma 5. *For any actions $a \neq a'$ and state $j \in S_a$:*

(a) $F_p(S) = F_p(\mathcal{T}_j^{a,a'} S) + f_j^{a,a'}(S) x_{pj}^{a'}(\mathcal{T}_j^{a,a'} S)$;

(b) $F_p(\mathcal{T}_j^{a,a'} S) = F_p(S) + f_j^{a',a}(\mathcal{T}_j^{a,a'} S) x_{pj}^{a}(S)$;

(c) $G_p(\mathcal{T}_j^{a,a'} S) = G_p(S) + g_j^{a,a'}(S) x_{pj}^{a'}(\mathcal{T}_j^{a,a'} S)$;

(d) $G_p(S) = G_p(\mathcal{T}_j^{a,a'} S) + g_j^{a',a}(\mathcal{T}_j^{a,a'} S) x_{pj}^{a}(S)$.

Proof. (a) Writing Lemma 4(a) as

$$F_p(S) = F_p(\pi) + \sum_{l \neq k} \sum_{j' \in S_l} f_{j'}^{l,k}(S) x_{pj'}^{k}(\pi)$$

and taking $\pi = \mathcal{T}_j^{a,a'} S$ in the latter expression, with $j \in S_a$, gives

$$F_p(S) = F_p(\mathcal{T}_j^{a,a'} S) + \sum_{l \neq k} \sum_{j' \in S_l} f_{j'}^{l,k}(S) x_{pj'}^{k}(\mathcal{T}_j^{a,a'} S). \tag{21}$$

Now, on the one hand, for $l \neq a$, policy $\mathcal{T}_j^{a,a'} S$ selects gear l in states $j' \in S_l$. Hence, for $k \neq l$ and such j', $x_{pj'}^{k}(\mathcal{T}_j^{a,a'} S) = 0$.

On the other hand, for $l = a$, policy $\pi = \mathcal{T}_j^{a,a'} S$ selects gear $l = a$ in states $j' \in S_l - \{j\}$ and selects gear a' in state $j' = j$. Hence, $x_{pj'}^{k}(\mathcal{T}_j^{a,a'} S) = 0$ for $j' \in S_l - \{j\}$, since $k \neq l = a$ and $x_{pj'}^{k}(\mathcal{T}_j^{a,a'} S) = 0$ for $j' = j$ if $k \neq a'$.

Thus, the only positive $x_{pj'}^{k}(\mathcal{T}_j^{a,a'} S)$ in (21) can be $x_{pj}^{a'}(\mathcal{T}_j^{a,a'} S)$; hence, (21) reduces to

$$F_p(S) = F_p(\mathcal{T}_j^{a,a'} S) + f_j^{a,a'}(S) x_{pj}^{a'}(\mathcal{T}_j^{a,a'} S).$$

(b) Let $S' = \mathcal{T}_j^{a,a'} S$ and note that $j \in S'_{a'}$ and $S = \mathcal{T}_j^{a',a} S'$. Hence, by part (a),

$$F_p(\mathcal{T}_j^{a,a'} S) = F_p(S') = F_p(\mathcal{T}_j^{a',a} S') + f_j^{a',a}(S') x_{pj}^{a}(\mathcal{T}_j^{a',a} S')$$
$$= F_p(S) + f_j^{a',a}(\mathcal{T}_j^{a,a'} S) x_{pj}^{a}(S).$$

Parts (c) and (d) follow similarly as (a) and (b). ☐

The following result, which follows easily from Lemma 5(c, d), clarifies the interpretation of PCL($\mathcal{F}$)-indexability condition (PCLI1) in Definition 3, as a natural monotonicity property of the resource metric $G_i(S)$.

Proposition 1. *Condition (PCLI1) is equivalent to the following: for $S \in \mathcal{F}, a' < a < a'', j \in S_a$,*

(a) $G_i(\mathcal{T}_j^{a,a'} S) \leqslant G_i(S) \leqslant G_i(\mathcal{T}_j^{a,a''} S)$, $i \neq j$, and $G_j(\mathcal{T}_j^{a,a'} S) < G_j(S) < G_j(\mathcal{T}_j^{a,a''} S)$.

(b)   *If p has full support, $G_p(\mathcal{T}_j^{a,a'} S) < G_p(S) < G_p(\mathcal{T}_j^{a,a''} S)$.*

Remark 4. *Proposition 1 yields the following intuitive interpretation of condition (PCLI1): it means that, for any $\mathcal{F}$-policy S, modifying S by downshifting gears in one state results in a lower or equal value of the resource usage metric; and, conversely, modifying S by upshifting gears in*

one state results in a higher or equal resource usage metric. When the initial state is drawn from a distribution with full support, downshifting gears leads to a strictly lower resource usage, and upshifting gears leads to a strictly higher resource usage.

The next result shows that, under (PCLI1), the increment $F_p(\mathcal{T}_j^{a,a'}S) - F_p(S)$ is proportional to $G_p(S) - G_p(\mathcal{T}_j^{a,a'}S)$, with proportionality constant $m_j^{a',a}(S)$, which is equal to $m_j^{a',a}(\mathcal{T}_j^{a,a'}S)$.

Lemma 6. *Under condition* (PCLI1), *for any actions* $a' \neq a$ *and state* $j \in S_a$,

(a) $F_p(\mathcal{T}_j^{a,a'}S) - F_p(S) = m_j^{a',a}(S)(G_p(S) - G_p(\mathcal{T}_j^{a,a'}S)) = m_j^{a',a}(\mathcal{T}_j^{a,a'}S)(G_p(S) - G_p(\mathcal{T}_j^{a,a'}S))$;

(b) $m_j^{a',a}(S) = m_j^{a',a}(\mathcal{T}_j^{a,a'}S)$.

Proof. (a) We have, using Lemma 5(a, c),

$$F_p(\mathcal{T}_j^{a,a'}S) - F_p(S) = -f_j^{a,a'}(S)x_{pj}^{a'}(\mathcal{T}_j^{a,a'}S) = -f_j^{a,a'}(S)\frac{G_p(\mathcal{T}_j^{a,a'}S) - G_p(S)}{g_j^{a,a'}(S)}$$

$$= m_j^{a',a}(S)(G_p(S) - G_p(\mathcal{T}_j^{a,a'}S)).$$

On the other hand, using Lemma 5(b, d) we obtain

$$F_p(\mathcal{T}_j^{a,a'}S) - F_p(S) = f_j^{a',a}(\mathcal{T}_j^{a,a'}S)x_{pj}^a(S) = f_j^{a',a}(\mathcal{T}_j^{a,a'}S)\frac{G_p(S) - G_p(\mathcal{T}_j^{a,a'}S))}{g_j^{a',a}(\mathcal{T}_j^{a,a'}S)}$$

$$= m_j^{a',a}(\mathcal{T}_j^{a,a'}S)(G_p(S) - G_p(\mathcal{T}_j^{a,a'}S)).$$

(b) This part follows directly from part (a). □

The following result shows in its part (a) that, under condition (PCLI1), the increment $f_j^{a',a}(S) - f_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S)$ is proportional to $g_j^{a',a}(S) - g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S)$, with the proportionality constant being $m_i^{\bar{a}',\bar{a}}(S)$. Then, its part (b) draws on this result to obtain a relation between MP metrics that we use in Algorithm 1.

Lemma 7. *Under condition* (PCLI1), *for any actions* $a' \neq a$ *and* $\bar{a}' \neq \bar{a}$ *with* $\mathcal{T}_i^{\bar{a},\bar{a}'}S \in \mathcal{F}$ *and states* $i \in S_{\bar{a}}$ *and* j,

(a) $f_j^{a',a}(S) - f_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S) = m_i^{\bar{a}',\bar{a}}(S)(g_j^{a',a}(S) - g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S))$;

(b) $g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S)(m_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S) - m_i^{\bar{a}',\bar{a}}(S)) = g_j^{a',a}(S)(m_j^{a',a}(S) - m_i^{\bar{a}',\bar{a}}(S))$.

Proof. (a) From (11) and (12), we obtain

$$\mathbf{f}^{a',a}(S) - \mathbf{f}^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S) = \beta(\mathbf{P}^{a'} - \mathbf{P}^a)(\mathbf{F}(S) - \mathbf{F}(\mathcal{T}_i^{\bar{a},\bar{a}'}S))$$

and

$$\mathbf{g}^{a',a}(S) - \mathbf{g}^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S) = \beta(\mathbf{P}^a - \mathbf{P}^{a'})(\mathbf{G}(S) - \mathbf{G}(\mathcal{T}_i^{\bar{a},\bar{a}'}S))$$

Now, combining Lemma 6(a) with this gives

$$\mathbf{f}^{a',a}(S) - \mathbf{f}^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S) = m_i^{\bar{a}',\bar{a}}(S)(\mathbf{g}^{a',a}(S) - \mathbf{g}^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'}S)).$$

(b) We can write

$$
\begin{aligned}
m_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S) &\triangleq \frac{f_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} \\
&= \frac{f_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} - m_i^{\bar{a}',\bar{a}}(S)\left(\frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} - 1\right) \\
&= \frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} \frac{f_j^{a',a}(S)}{g_j^{a',a}(S)} - m_i^{\bar{a}',\bar{a}}(S)\left(\frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} - 1\right) \\
&= \frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} m_j^{a',a}(S) - m_i^{\bar{a}',\bar{a}}(S)\left(\frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)} - 1\right) \\
&= m_i^{\bar{a}',\bar{a}}(S) + \frac{g_j^{a',a}(S)}{g_j^{a',a}(\mathcal{T}_i^{\bar{a},\bar{a}'} S)}\left(m_j^{a',a}(S) - m_i^{\bar{a}',\bar{a}}(S)\right),
\end{aligned}
$$

where the second line is obtained by using part (a). □

5. Analysis of the Output of Algorithm DS($\mathcal{F}$)

This section derives further relations between project performance metrics that will play the role of key tools for elucidating and analyzing the output of the downshift adaptive-greedy index algorithm DS($\mathcal{F}$) in Algorithm 1 and hence to prove the verification theorem. Throughout this section, $\{(j_k, a_k), S^k, m_{j_k}^{*,a_k}\}_{k=1}^K$ is an output of algorithm DS($\mathcal{F}$).

We start with a result justifying recursive index update Formula (14) in the algorithm.

Lemma 8. *Let condition* (PCLI1) *hold. Then, for any state j, active action $a \geqslant 1$, and $k = 2, \ldots, K$,*

$$
m_j^{a-1,a}(S^k) = m_{j_{k-1}}^{*,a_{k-1}} + \frac{g_j^{a-1,a}(S^{k-1})}{g_j^{a-1,a}(S^k)}\left(m_j^{a-1,a}(S^{k-1}) - m_{j_{k-1}}^{*,a_{k-1}}\right), \tag{22}
$$

or, equivalently,

$$
m_j^{a-1,a}(S^{k-1}) = m_{j_{k-1}}^{*,a_{k-1}} + \frac{g_j^{a-1,a}(S^k)}{g_j^{a-1,a}(S^{k-1})}\left(m_j^{a-1,a}(S^k) - m_{j_{k-1}}^{*,a_{k-1}}\right). \tag{23}
$$

Proof. The result follows by noting that $m_{j_{k-1}}^{*,a_{k-1}} = m_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}$ and taking $i = j_{k-1}$, $\bar{a}' = a_{k-1} - 1$, $\bar{a} = a_{k-1}$, $S = S^{k-1}$, and $a' = a - 1$ in Lemma 7(b), since $\mathcal{T}_i^{\bar{a},\bar{a}'} S = S^k \in \mathcal{F}$. □

The following result expresses, in its part (b), the MP metric $m_{j_l}^{a_l-1,a_l}(S^k)$, for $k < l$, as the sum of index $m_{j_k}^{*,a_k}$ and a positive linear combination of the index differences $m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}$ for $n = k+1, \ldots, l$. Part (a) is a preliminary result needed to prove part (b).

Lemma 9. *Under condition* (PCLI1), *the following holds for $1 \leqslant k < l \leqslant K$:*
(a)   For $l' = k+1, \ldots, l$,

$$
\begin{aligned}
m_{j_l}^{a_l-1,a_l}(S^k) = m_{j_k}^{*,a_k} + \frac{1}{g_{j_l}^{a_l-1,a_l}(S^k)}\Bigg[&\sum_{n=k+1}^{l'-1} g_{j_l}^{a_l-1,a_l}(S^n)\left(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}\right) \\
&+ g_{j_l}^{a_l-1,a_l}(S^{l'})\left(m_{j_l}^{a_l-1,a_l}(S^{l'}) - m_{j_{l'-1}}^{*,a_{l'-1}}\right)\Bigg].
\end{aligned} \tag{24}
$$

(b) $f_{j_l}^{a_l-1,a_l}(S^k) = g_{j_l}^{a_l-1,a_l}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l} g_{j_l}^{a_l-1,a_l}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$, *or, equivalently,*

$$m_{j_l}^{a_l-1,a_l}(S^k) = m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l} \frac{g_{j_l}^{a_l-1,a_l}(S^n)}{g_{j_l}^{a_l-1,a_l}(S^k)}(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}). \tag{25}$$

Proof. (a) We prove the result by induction on $l' = k+1, \ldots, l$. For $l' = k+1$, Equation (24) holds because, by Equation (23) in Lemma 8, we have

$$m_{j_l}^{a_l-1,a_l}(S^k) = m_{j_k}^{*,a_k} + \frac{g_{j_l}^{a_l-1,a_l}(S^{k+1})}{g_{j_l}^{a_l-1,a_l}(S^k)}(m_{j_l}^{a_l-1,a_l}(S^{k+1}) - m_{j_k}^{*,a_k}).$$

Suppose now that Equation (24) holds for some l' with $k < l' < l$. We will use that, by Equation (23) in Lemma 8, we have

$$m_{j_l}^{a_l-1,a_l}(S^{l'}) = m_{j_{l'}}^{*,a_{l'}} + \frac{g_{j_l}^{a_l-1,a_l}(S^{l'+1})}{g_{j_l}^{a_l-1,a_l}(S^{l'})}(m_{j_l}^{a_l-1,a_l}(S^{l'+1}) - m_{j_{l'}}^{*,a_{l'}}).$$

Now, substituting the right-hand side of the latter identity for $m_{j_l}^{a_l-1,a_l}(S^{l'})$ in Equation (24) gives

$$m_{j_l}^{a_l-1,a_l}(S^k) = m_{j_k}^{*,a_k} + \frac{1}{g_{j_l}^{a_l-1,a_l}(S^k)}\left[\sum_{n=k+1}^{l'-1} g_{j_l}^{a_l-1,a_l}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}) \right.$$

$$\left. + g_{j_l}^{a_l-1,a_l}(S^{l'})\left(m_{j_{l'}}^{*,a_{l'}} + \frac{g_{j_l}^{a_l-1,a_l}(S^{l'+1})}{g_{j_l}^{a_l-1,a_l}(S^{l'})}(m_{j_l}^{a_l-1,a_l}(S^{l'+1}) - m_{j_{l'}}^{*,a_{l'}}) - m_{j_{l'-1}}^{*,a_{l'-1}} \right)\right]$$

$$= m_{j_k}^{*,a_k} + \frac{1}{g_{j_l}^{a_l-1,a_l}(S^k)}\left[\sum_{n=k+1}^{l'} g_{j_l}^{a_l-1,a_l}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}) \right.$$

$$\left. + g_{j_l}^{a_l-1,a_l}(S^{l'+1})(m_{j_l}^{a_l-1,a_l}(S^{l'+1}) - m_{j_{l'}}^{*,a_{l'}}) \right],$$

which shows that the result also holds for $l' + 1$ and hence completes the induction.

(b) This part corresponds to the case $l' = l$ in part (a), noting that $m_{j_l}^{a_l-1,a_l}(S^l) = m_{j_l}^{*,a_l}$. $\square$

In the following result and henceforth we use the notation $a^k(j)$ to denote the action selected in state j by policy S^k. Thus, e.g., $a^1(j) = A$ for every state j, since $S^1 = (\emptyset, \ldots, \emptyset, \mathcal{N})$, and $a^{K+1}(j) = 0$ for every state j, since $S^{K+1} = (\mathcal{N}, \emptyset, \ldots, \emptyset)$.

Lemma 10. *Under condition* (PCLI1), *the following holds for* $1 \leqslant k \leqslant l \leqslant K$:

$$f_{j_l}^{a_l-1,a^k(j_l)}(S^k) = g_{j_l}^{a_l-1,a^k(j_l)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l} g_{j_l}^{a_l-1,a^n(j_l)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}), \tag{26}$$

or, equivalently,

$$m_{j_l}^{a_l-1,a^k(j_l)}(S^k) = m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l} \frac{g_{j_l}^{a_l-1,a^n(j_l)}(S^n)}{g_{j_l}^{a_l-1,a^k(j_l)}(S^k)}(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}).$$

Proof. Fix k. We prove the result by induction on $l = k, \ldots, K$. For $l = k$, using that $a^k(j_k) = a_k$ gives that (26) reduces to $f_{j_l}^{a_k-1,a_k}(S^k) = g_{j_k}^{a_k-1,a_k}(S^k)m_{j_k}^{*,a_k}$, which holds by construction, since $m_{j_k}^{*,a_k}$ is defined in the algorithm precisely as $m_{j_k}^{a_k-1,a_k}(S^k) = f_{j_l}^{a_k-1,a_k}(S^k)/g_{j_k}^{a_k-1,a_k}(S^k)$.

Suppose now that (26) holds up to and including some l with $k < l < K$. We will prove that it must then hold for $l + 1$. For such a purpose, we distinguish two cases, depending on whether $j_{l+1} = j_l$ or $j_{l+1} \neq j_l$. Start with the case $j_{l+1} = j_l$. To simplify the argument below, we write j_l as j and a_l as a. In this case, the algorithm downshifts in step l the gear in state j from a to $a - 1$ and in step $l + 1$ downshifts again in state j from gear $a - 1$ to $a - 2$. Hence, $a^l(j) = a$ and $a^{l+1}(j) = a_{l+1} = a - 1$. We can write

$$f_{j_{l+1}}^{a_{l+1}-1,a^k(j_{l+1})}(S^k) = f_{j_{l+1}}^{a_{l+1}-1,a_{l+1}}(S^k) + f_{j_{l+1}}^{a_{l+1},a^k(j)}(S^k) = f_j^{a-2,a-1}(S^k) + f_j^{a-1,a^k(j)}(S^k)$$

$$= g_j^{a-2,a-1}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_j^{a-2,a-1}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$+ g_j^{a-1,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l} g_j^{a-1,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$= g_j^{a-2,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_j^{a-2,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$= g_{j_{l+1}}^{a_{l+1}-1,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_{j_{l+1}}^{a_{l+1}-1,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}),$$

where we have used in turn the elementary property $f_j^{a,a''}(S) = f_j^{a,a'}(S) + f_j^{a',a''}(S)$, Lemma 9(b), the induction hypothesis, and $a^{l+1}(j) = a - 1$. Therefore, the result holds for $l + 1$ in this case.

Consider now the case $j_{l+1} \neq j_l$, in which $a_{l+1} = a^{l+1}(j_{l+1}) = a^l(j_{l+1})$. To simplify the argument below, we write j_{l+1} as j and a_{l+1} as a. In this case, the algorithm downshifts in step $l + 1$ at state j from gear a to $a - 1$. If $a < A$, the previous downshift at state j, from gear $a + 1$ to a, occurred at some earlier step $l' < l$, so $j_{l'} = j$, $a_{l'} - 1 = a$, and

$$a^{l'+1}(j) = \cdots = a^l(j) = a^{l+1}(j) = a. \tag{27}$$

We can now write

$$f_{j_{l+1}}^{a_{l+1}-1,a^k(j_{l+1})}(S^k) = f_{j_{l+1}}^{a_{l+1}-1,a_{l+1}}(S^k) + f_{j_{l+1}}^{a_{l+1},a^k(j)}(S^k) = f_j^{a-1,a}(S^k) + f_j^{a,a^k(j)}(S^k)$$

$$= g_j^{a-1,a}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_j^{a-1,a}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$+ g_{j_{l'}}^{a_{l'}-1,a^k(j_{l'})}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l'} g_{j_l}^{a_{l'}-1,a^n(j_{l'})}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$= g_j^{a-1,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l'} g_j^{a-1,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$+ \sum_{n=l'+1}^{l+1} g_j^{a-1,a}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$= g_{j_{l+1}}^{a_{l+1}-1,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_{j_{l+1}}^{a_{l+1}-1,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}),$$

where we have used in turn Lemma 9(b), the induction hypothesis, and (27).

Suppose now that $a = A$. Then,

$$a^k(j) = a^{k+1}(j) = \cdots = a^{l+1}(j) = A, \tag{28}$$

and we can write

$$f_{j_{l+1}}^{a_{l+1}-1,a^k(j_{l+1})}(S^k) = f_j^{A-1,A}(S^k)$$

$$= g_j^{A-1,A}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_j^{A-1,A}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}})$$

$$= g_{j_{l+1}}^{a_{l+1}-1,a^k(j)}(S^k)m_{j_k}^{*,a_k} + \sum_{n=k+1}^{l+1} g_{j_{l+1}}^{a_{l+1}-1,a^n(j)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}),$$

where we have used in turn Lemma 9(b) and (28).

Hence, the result also holds for $l+1$ in the case $j_{l+1} \neq j_l$, which completes the induction proof. $\square$

The next result expresses the modified holding costs $\widehat{h}_j^a$ as positive linear combinations of $m_{j_1}^{*,a_1}$ and the differences $m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}$. We will use it in Lemmas 14 and 15 to reformulate the modified cost objective $\widehat{V}_p(\lambda, \pi)$.

Lemma 11. *Under condition (PCLI1),*

$$g_{j_1}^{a_1-1,A}(S^1)m_{j_1}^{*,a_1} = \widehat{h}_{j_1}^{a_1-1}$$

$$g_{j_l}^{a_l-1,A}(S^1)m_{j_1}^{*,a_1} + \sum_{n=2}^{l} g_{j_l}^{a_l-1,a^n(j_l)}(S^n)(m_{j_n}^{*,a_n} - m_{j_{n-1}}^{*,a_{n-1}}) = \widehat{h}_{j_l}^{a_l-1}, \quad l = 2, \ldots, K. \tag{29}$$

Proof. The result follows from Corollary 1(a), which shows that $f_j^{a-1,A}(S^1) = \widehat{h}_j^{a-1} - \widehat{h}_j^A = \widehat{h}_j^{a-1}$ (since $\widehat{h}_j^A = 0$), and Lemma 10, used with $k = 1$, noting that $a^1(j) = A$. $\square$

The following result shows that $m_{j_{k-1}}^{*,a_{k-1}}$ can be expressed in two different ways in terms of MP metrics.

Lemma 12. *Under condition (PCLI1),*

$$m_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k) = m_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^{k-1}) = m_{j_{k-1}}^{*,a_{k-1}}, \quad k = 2, \ldots, K+1.$$

Proof. The first identity follows from the result $m_j^{a',a}(\mathcal{T}_j^{a,a'}S) = m_j^{a',a}(S)$ in Lemma 6(b), taking $S = S^{k-1}$, $a = a_{k-1}$, $a' = a_{k-1} - 1$, and $j = j_{k-1}$, and noting that $S^k = \mathcal{T}_{j_{k-1}}^{a_{k-1},a_{k-1}-1}S^{k-1}$. The second identity follows by definition of $m_{j_{k-1}}^{*,a_{k-1}}$. $\square$

The following result relates metrics under two successive policies as generated by the index algorithm.

Lemma 13. *Under condition (PCLI1),*
(a) $F_p(S^k) = F_p(S^{k-1}) + f_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}), \quad k = 2, \ldots, K+1;$
(b) $F_p(S^k) = F_p(S^{k+1}) - f_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k), \quad k = 1, \ldots, K;$
(c) $G_p(S^k) = G_p(S^{k-1}) - g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}), \quad k = 2, \ldots, K+1;$
(d) $G_p(S^k) = G_p(S^{k+1}) + g_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k), \quad k = 1, \ldots, K;$
(e) $V_p(\lambda, S^k) = V_p(\lambda, S^{k-1}) - (\lambda - m_{j_{k-1}}^{*,a_{k-1}})g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}), \quad k = 2, \ldots, K+1.$
(f) $V_p(\lambda, S^k) = V_p(\lambda, S^{k+1}) - (m_{j_k}^{*,a_k} - \lambda)g_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k), \quad k = 1, \ldots, K.$

Proof. (a) This part follows from Lemma 5(a) by taking $S = S^k$, $a = a_{k-1} - 1$, $a' = a_{k-1}$, and $j = j_{k-1}$, noting that $\mathcal{T}_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}S^k = S^{k-1}$ in $F_p(S) = F_p(\mathcal{T}_j^{a,a'}S) + f_j^{a,a'}(S)x_{pj}^{a'}(\mathcal{T}_j^{a,a'}S)$.

(b) This part follows directly from (a).

(c) This part follows from Lemma 5(c) by taking S, a, a', and j as in part (a) in $G_p(\mathcal{T}_j^{a,a'}S) = G_p(S) + g_j^{a,a'}(S)x_{pj}^{a'}(\mathcal{T}_j^{a,a'}S)$.

(d) This part follows directly from (c).

(e) The result follows from

$$
\begin{aligned}
V_p(\lambda, S^k) &= F_p(S^k) + \lambda G_p(S^k) \\
&= F_p(S^{k-1}) + f_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}) \\
&\quad + \lambda\big[G_p(S^{k-1}) - g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1})\big] \\
&= V_p(\lambda, S^{k-1}) + \big(f_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k) - \lambda g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)\big)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}) \\
&= V_p(\lambda, S^{k-1}) + \big(m_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k) - \lambda\big)g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}) \\
&= V_p(\lambda, S^{k-1}) - \big(\lambda - m_{j_{k-1}}^{*,a_{k-1}}\big)g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}),
\end{aligned}
$$

where we have used parts (a, c) and Lemma 12.

(f) The result follows from

$$
\begin{aligned}
V_p(\lambda, S^k) &= F_p(S^k) + \lambda G_p(S^k) \\
&= F_p(S^{k+1}) - f_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k) \\
&\quad + \lambda\big[G_p(S^{k+1}) + g_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k)\big] \\
&= V_p(\lambda, S^{k+1}) - \big(f_{j_k}^{a_k-1,a_k}(S^{k+1}) - \lambda g_{j_k}^{a_k-1,a_k}(S^{k+1})\big)x_{pj_k}^{a_k}(S^k) \\
&= V_p(\lambda, S^{k+1}) - \big(m_{j_k}^{a_k-1,a_k}(S^{k+1}) - \lambda\big)g_{j_k}^{a_k-1,a_k}(S^{k+1})x_{pj_k}^{a_k}(S^k) \\
&= V_p(\lambda, S^{k-1}) - \big(\lambda - m_{j_{k-1}}^{*,a_{k-1}}\big)g_{j_{k-1}}^{a_{k-1}-1,a_{k-1}}(S^k)x_{pj_{k-1}}^{a_{k-1}}(S^{k-1}),
\end{aligned}
$$

where we have used parts (b, d) and Lemma 12. $\quad\square$

6. Partial Conservation Laws

This section shows that, under condition (PCLI1) in Definition 3, project performance metrics satisfy certain *partial conservation laws* (PCLs), which extend those previously introduced by the author for finite-state restless (two-gear) bandits in [25,29]. It further uses those PCLs to lay further groundwork towards the proof of Theorem 1.

In the following result, we assume that the initial-state distribution p has full support, which we write as $p > 0$.

Proposition 2 (PCLs). *Suppose that* (PCLI1) *holds and let* $p > 0$. *Then, metrics* $G_p(\pi)$ *and* $x_{pj}^a(\pi)$, *for states j and actions* $a < A$, *satisfy the following: for any admissible policy* π *and* $S \in \mathcal{F}$,

(a.1) $G_p(\pi) + \sum_{a<a'}\sum_{j\in S_{a'}} g_j^{a,a'}(S)x_{pj}^a(\pi) \geqslant G_p(S)$, *with equality (conservation law),*

$$
G_p(\pi) + \sum_{a<a'}\sum_{j\in S_{a'}} g_j^{a,a'}(S)x_{pj}^a(\pi) = G_p(S), \tag{30}
$$

iff π *selects gears* $a' \leqslant a$ *in states* $j \in S_a$ *(so* $\pi \preceq S$), *for* $a = 0,\dots,A-1$.

(a.2) In particular, for $S = S^{K+1} = (\mathcal{N}, \varnothing, \dots, \varnothing)$, *it holds that* $G_p(\pi) \geqslant G_p(S^{K+1})$, *with equality iff* π *selects gear 0 in every state.*

(a.3) In the case $S = S^1 = (\varnothing, \dots, \varnothing, \mathcal{N})$, *we have the conservation law*

$$
G_p(\pi) + \sum_{a<A}\sum_{j\in\mathcal{N}} g_j^{a,A}(S^1)x_{pj}^a(\pi) = G_p(S^1). \tag{31}
$$

(b) $\sum_{a<a'}\sum_{j\in S_{a'}} g_j^{a,a'}(S)x_{pj}^a(\pi) \geqslant 0$, *with equality iff π selects gears $a \geqslant a'$ in states $j \in S_{a'}$ (so $S \preceq \pi$), for $a' = 1, \ldots, A$.*

Proof. (a.1) From Lemma 4(b), we obtain, under condition (PCLI1),

$$G_p(\pi) + \sum_{a<a'}\sum_{j\in S_{a'}} g_j^{a,a'}(S)x_{pj}^a(\pi) = G_p(S) + \sum_{a<a'}\sum_{j\in S_a} g_j^{a,a'}(S)x_{pj}^{a'}(\pi) \geqslant G_p(S), \qquad (32)$$

with equality iff (using the fact that the initial state distribution p has full support) $x_{pj}^{a'}(\pi) = 0$ for $j \in S_a$ with $a < a'$, i.e., iff π selects gears $a' \leqslant a$ in states $j \in S_a$, for $a = 0, \ldots, A-1$.

Parts (a.2) and (a.3) are direct consequences of (a.1).

Part (b) follows directly from (PCLI1).   □

The next result shows how to reformulate the equivalent modified holding cost metric $\widehat{F}_p(\pi) \triangleq \sum_{a=0}^{A-1}\sum_{j\in\mathcal{N}} \widehat{h}_j^a x_{pj}^a(\pi)$ (see Section 3) in terms of the output $\{(j_k, a_k), S^k, m_{j_k}^{*,a_k}\}_{k=1}^K$ of algorithm DS($\mathcal{F}$) in Algorithm 1 and expressions arising in the PCLs in Proposition 2.

Lemma 14. *Under condition (PCLI1),*

$$\widehat{F}_p(\pi) = m_{j_1}^{*,a_1} \sum_{a<a'}\sum_{j\in S_{a'}^1} g_j^{a,a'}(S^1)x_{pj}^a(\pi) + \sum_{k=2}^K (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a<a'}\sum_{j\in S_{a'}^k} g_j^{a,a'}(S^k)x_{pj}^a(\pi).$$

Proof. The result follows from Lemma 11, which yields

$$\begin{aligned}
\widehat{F}_p(\pi) &\triangleq \sum_{a=0}^{A-1}\sum_{j\in\mathcal{N}} \widehat{h}_j^a x_{pj}^a(\pi) = \sum_{l=1}^K \widehat{h}_{j_l}^{a_l-1} x_{pj_l}^{a_l-1}(\pi) \\
&= g_{j_1}^{a_1-1,A}(S^1)m_{j_1}^{*,a_1} x_{pj_1}^{a_1-1}(\pi) \\
&\quad + \sum_{l=2}^K \left[g_{j_l}^{a_l-1,A}(S^1)m_{j_1}^{*,a_1} + \sum_{k=2}^l g_{j_l}^{a_l-1,a^k(j_l)}(S^k)(m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \right] x_{pj_l}^{a_l-1}(\pi) \\
&= m_{j_1}^{*,a_1} \sum_{l=1}^K g_{j_l}^{a_l-1,A}(S^1)x_{pj_l}^{a_l-1}(\pi) + \sum_{k=2}^K (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{l=k}^K g_{j_l}^{a_l-1,a^k(j_l)}(S^k)x_{pj_l}^{a_l-1}(\pi) \\
&= m_{j_1}^{*,a_1} \sum_{a<a'}\sum_{j\in S_{a'}^1} g_j^{a,a'}(S^1)x_{pj}^a(\pi) + \sum_{k=2}^K (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a<a'}\sum_{j\in S_{a'}^k} g_j^{a,a'}(S^k)x_{pj}^a(\pi).
\end{aligned}$$

□

The following result draws on the previous one by showing how to reformulate the cost metric $\widehat{V}_p(\lambda, \pi)$ of the equivalent modified λ-price problem (see Lemma 3) in terms of the output of algorithm DS($\mathcal{F}$).

Lemma 15. *Under (PCLI1), $\widehat{V}_p(\lambda, \pi)$ can be reformulated into the following equivalent expressions:*
(a)

$$\widehat{V}_p(\lambda, \pi) = \lambda G_p(S^1) + (m_{j_1}^{*,a_1} - \lambda) \sum_{a<a'}\sum_{j\in S_{a'}^1} g_j^{a,a'}(S^1)x_{pj}^a(\pi)$$

$$+ \sum_{k=2}^K (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a<a'}\sum_{j\in S_{a'}^k} g_j^{a,a'}(S^k)x_{pj}^a(\pi);$$

(b)

$$\widehat{V}_p(\lambda, \pi) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{l-1} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \left(G_p(\pi) + \sum_{a<a'} \sum_{j \in S_{a'}^k} g_j^{a,a'}(S^k) x_{pj}^a(\pi) \right)$$

$$+ (\lambda - m_{j_{l-1}}^{*,a_{l-1}}) \left(G_p(\pi) + \sum_{a<a'} \sum_{j \in S_{a'}^l} g_j^{a,a'}(S^l) x_{pj}^a(\pi) \right)$$

$$+ (m_{j_l}^{*,a_l} - \lambda) \sum_{a<a'} \sum_{j \in S_{a'}^l} g_j^{a,a'}(S^l) x_{pj}^a(\pi)$$

$$+ \sum_{k=l+1}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a<a'} \sum_{j \in S_{a'}^k} g_j^{a,a'}(S^k) x_{pj}^a(\pi);$$

(c)

$$\widehat{V}_p(\lambda, \pi) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \left(G_p(\pi) + \sum_{a<a'} \sum_{j \in S_{a'}^k} g_j^{a,a'}(S^k) x_{pj}^a(\pi) \right)$$

$$+ (\lambda - m_{j_K, a_K}) G_p(\pi).$$

Proof. All three parts follow from the definition $\widehat{V}_p(\lambda, \pi) \triangleq \widehat{F}_p(\pi) + \lambda G_p(\pi)$ and Lemma 14 by suitably rearranging terms. □

The next result draws on the above to reformulate the cost metrics $\widehat{V}_p(\lambda, S^l)$ in terms of the output of algorithm DS($\mathcal{F}$).

Lemma 16. *Under condition (PCLI1),*
(a) $\widehat{V}_p(\lambda, S^1) = \lambda G_p(S^1)$;
(b) *For $2 \leqslant l \leqslant K$,*

$$\widehat{V}_p(\lambda, S^l) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{l-1} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) G_p(S^k) + (\lambda - m_{j_{l-1}}^{*,a_{l-1}}) G_p(S^l);$$

(c) $\widehat{V}_p(\lambda, S^{K+1}) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) G_p(S^k) + (\lambda - m_{j_K}^{*,a_K}) G_p(S^{K+1})$.

Proof. Each part follows from the corresponding part in Lemma 15, using the PCLs in Proposition 2 to simplify the resulting expressions. □

7. Proof of the Verification Theorem

We are now ready to prove Theorem 1.

Proof of Theorem 1. We will prove the result by showing the following: (i) policy S^1 is λ-optimal iff $\lambda \leqslant m_{j_1}^{*,a_1}$; (ii) for $2 \leqslant l \leqslant K$, policy S^l is λ-optimal iff $m_{j_{l-1}}^{*,a_{l-1}} \leqslant \lambda \leqslant m_{j_l}^{*,a_l}$; and (iii) policy S^{K+1} is λ-optimal iff $\lambda \geqslant m_{j_K}^{*,a_K}$. Note that (i, ii, iii) imply that the model is $\mathcal{F}$-indexable with DAI $\lambda_j^{*,a}$ being given by the MPI $m_j^{*,a}$.

We consider below that $p > 0$, i.e., the initial-state distribution p has full support.

Start with (i). If $\lambda \leqslant m_{j_1}^{*,a_1}$, we have, for any policy π,

$$\widehat{V}_p(\lambda, \pi) = \lambda G_p(S^1) + (m_{j_1}^{*,a_1} - \lambda) \sum_{a < a'} \sum_{j \in S^1_{a'}} g_j^{a,a'}(S^1) x_{pj}^a(\pi)$$

$$+ \sum_{k=2}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a < a'} \sum_{j \in S^k_{a'}} g_j^{a,a'}(S^k) x_{pj}^a(\pi)$$

$$\geqslant \lambda G_p(S^1) = \widehat{V}_p(\lambda, S^1),$$

where we have used Lemmas 15(a) and 16(a) and conditions (PCLI1, PCLI2). Hence, policy S^1 is λ-optimal.

Conversely, suppose that policy S^1 is λ-optimal. Then, using Lemma 13(f), we obtain

$$0 \leqslant V_p(\lambda, S^2) - V_p(\lambda, S^1) = (m_{j_1}^{*,a_1} - \lambda) g_{j_1}^{a_1-1,a_1}(S^2) x_{pj_1}^{a_1}(S^1).$$

Now, since $g_{j_1}^{a_1-1,a_1}(S^2) > 0$ (by (PCLI1)) and $x_{pj_1}^{a_1}(S^1) > 0$ (because p has full support), it follows that $\lambda \leqslant m_{j_1}^{*,a_1}$.

Consider now (ii). If $m_{j_{l-1}}^{*,a_{l-1}} \leqslant \lambda \leqslant m_{j_l}^{*,a_l}$ for some l with $2 \leqslant l \leqslant K$, we have, for any policy π,

$$\widehat{V}_p(\lambda, \pi) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{l-1} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \left(G_p(\pi) + \sum_{a < a'} \sum_{j \in S^k_{a'}} g_j^{a,a'}(S^k) x_{pj}^a(\pi) \right)$$

$$+ (\lambda - m_{j_{l-1}}^{*,a_{l-1}}) \left(G_p(\pi) + \sum_{a < a'} \sum_{j \in S^l_{a'}} g_j^{a,a'}(S^l) x_{pj}^a(\pi) \right)$$

$$+ (m_{j_l}^{*,a_l} - \lambda) \sum_{a < a'} \sum_{j \in S^l_{a'}} g_j^{a,a'}(S^l) x_{pj}^a(\pi)$$

$$+ \sum_{k=l+1}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \sum_{a < a'} \sum_{j \in S^k_{a'}} g_j^{a,a'}(S^k) x_{pj}^a(\pi)$$

$$\geqslant \widehat{V}_p(\lambda, S^l),$$

where we have further used Lemmas 15(b) and 16(b), Proposition 2, and conditions (PCLI1, PCLI2). Hence, policy S^l is λ-optimal.

Conversely, suppose that policy S^l is λ-optimal. Then, using Lemma 13(e, f), we obtain

$$0 \leqslant V_p(\lambda, S^{l+1}) - V_p(\lambda, S^l) = (m_{j_l}^{*,a_l} - \lambda) g_{j_l}^{a_l-1,a_l}(S^{l+1}) x_{pj_l}^{a_l}(S^l)$$

and

$$0 \leqslant V_p(\lambda, S^{l-1}) - V_p(\lambda, S^l) = (\lambda - m_{j_{l-1}}^{*,a_{l-1}}) g_{j_{l-1}}^{a_{l-1}-1,a_{l-1}}(S^l) x_{pj_{l-1}}^{a_{l-1}}(S^{l-1}).$$

Now, since $g_{j_l}^{a_l-1,a_l}(S^{l+1}) > 0$, $g_{j_{l-1}}^{a_{l-1}-1,a_{l-1}}(S^l) > 0$, $x_{pj_l}^{a_l}(S^l) > 0$ and $x_{pj_{l-1}}^{a_{l-1}}(S^{l-1}) > 0$, it follows that $m_{j_{l-1}}^{*,a_{l-1}} \leqslant \lambda \leqslant m_{j_l}^{*,a_l}$.

Finally, consider (iii). If $\lambda \geqslant m_{j_K}^{*,a_K}$, we can write, for any policy π,

$$\widehat{V}_p(\lambda, \pi) = m_{j_1}^{*,a_1} G_p(S^1) + \sum_{k=2}^{K} (m_{j_k}^{*,a_k} - m_{j_{k-1}}^{*,a_{k-1}}) \left(G_p(\pi) + \sum_{a < a'} \sum_{j \in S^k_{a'}} g_j^{a,a'}(S^k) x_{pj}^a(\pi) \right)$$

$$+ (\lambda - m_{j_K,a_K}) G_p(\pi)$$

$$\geqslant \widehat{V}_p(\lambda, S^{K+1}),$$

where we have further used Lemmas 15(c) and 16(c), Proposition 2, and conditions (PCLI1, PCLI2). Hence, policy S^{K+1} is λ-optimal.

Conversely, suppose that policy S^{K+1} is λ-optimal. Then, using Lemma 13(f), we obtain

$$0 \leqslant V_p(\lambda, S^K) - V_p(\lambda, S^{K+1}l) = (\lambda - m_{j_K}^{*,a_K}) g_{j_K}^{a_K-1,a_K}(S^{K+1}) x_{pj_K}^{a_K}(S^K).$$

Now, since $g_{j_K}^{a_K-1,a_K}(S^{K+1}) > 0$ and $x_{pj_K}^{a_K}(S^K) > 0$, it follows that $\lambda \geqslant m_{j_K}^{*,a_K}$. This completes the proof. $\square$

8. Application to Multi-Armed Multi-Gear Bandit Problem: Bound and Index Policy

8.1. The Multi-Armed Multi-Gear Bandit Problem (MAMGBP)

Besides the intrinsic interest of the indexability property in Definition 1 for solving optimally the multi-gear bandit model, we next discuss as further motivation for such a property its application to design a suboptimal heuristic policy for the intractable *multi-armed multi-gear bandit problem* (MAMGBP) introduced by the author in [35] (where it was called the *multi-armed multi-mode bandit problem*).

The MAMGBP concerns the optimal dynamic allocation of a single shared resource to a finite collection of L projects modeled as multi-gear bandits, subject to a *peak resource constraint* stating that the total resource usage in each period cannot exceed a given amount $\bar{q}$. Denote by $s_l(t)$ and $a_l(t)$ the state and the action at time t for project $l = 1, \ldots, L$, which belong to the state and action spaces $\mathcal{N}_l = \{1, \ldots, N_l\}$ and $\mathcal{A}_l = \{0, \ldots, A_l\}$, respectively. The parameters of project l are denoted here by $h_l(j_l, a_l)$, $q_l(j_l, a_l)$, and $p_l^a(i_l, j_l)$.

The MAMGBP is a multi-dimensional MDP with *joint state* $\mathbf{s}(t) = (s_l(t))_{l=1}^L$ belonging to the *joint state space* $\mathcal{N} \triangleq \prod_{l=1}^L \mathcal{N}_l$ and *joint action* $\mathbf{a}(t) = (a_l(t))_{l=1}^L$.

The *joint holding cost* and *joint resource consumption* are additive across projects, being $h(\mathbf{i}, \mathbf{a}) \triangleq \sum_{l=1}^L h_l(i_l, a_l)$ and $q(\mathbf{i}, \mathbf{a}) \triangleq \sum_{l=1}^L q_l(i_l, a_l)$ in joint state $\mathbf{i} = (i_l)_{l=1}^L$ under joint action $\mathbf{a} = (a_l)_{l=1}^L$. The set of *feasible actions* in joint state $\mathbf{i}$, satisfying the aforementioned peak resource constraint, is

$$\mathcal{A}(\mathbf{i}) \triangleq \left\{ \mathbf{a} \in \prod_{l=1}^L \mathcal{A}_l \colon q(\mathbf{i}, \mathbf{a}) \leqslant \bar{q} \right\}. \tag{33}$$

To ensure that there always exists a feasible joint action, we require that, for every joint state $\mathbf{i}$,

$$q(\mathbf{i}, \mathbf{0}) \leqslant \bar{q}. \tag{34}$$

Individual project state transitions are conditionally independent given that the actions at every project have been selected, and hence the *joint transition probabilities* are multiplicative across projects, being given by $p_l^{\mathbf{a}}(\mathbf{i}, \mathbf{j}) \triangleq \prod_{l=1}^L p_l^{a_l}(i_l, j_l)$.

Let $\Pi(\bar{q})$ be the class of history-dependent randomized policies for selecting a feasible joint action at each time period, where we make explicit its dependence on $\bar{q}$ and denote by $\mathbb{E}_{\mathbf{i}}^{\pi}[\cdot]$ the expectation under policy $\pi \in \Pi(\bar{q})$ starting from the joint state $\mathbf{i}$. The expected total discounted holding cost incurred under policy π starting from $\mathbf{i}$ is

$$F(\mathbf{i}, \pi) \triangleq \mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{l=1}^L \sum_{t=0}^\infty h_l(s_l(t), a_l(t)) \beta^t \right],$$

and hence the optimal holding cost is

$$F^*(\mathbf{i}) \triangleq \inf \{ F(\mathbf{i}, \pi) \colon \pi \in \Pi(\bar{q}) \}.$$

We can thus formulate the MAMGBP as follows:

$$(P) \quad \text{find } \pi^* \in \Pi(\bar{q}) \colon F(\mathbf{i}, \pi^*) = F^*(\mathbf{i}), \quad \mathbf{i} \in \mathcal{N}. \tag{35}$$

We shall refer to a policy π^* solving the MAMGBP (P) as a *P-optimal policy*.

Again, standard results in MDP theory ensure the existence of a *P*-optimal policy π^* in the class $\mathbf{\Pi}^{\mathrm{SD}}$ of stationary deterministic policies, which is determined by the Bellman equations

$$F^*(\mathbf{i}) = \min_{\mathbf{a} \in \mathcal{A}(\mathbf{i})} h(\mathbf{i}, \mathbf{a}) + \beta \sum_{\mathbf{j} \in \mathcal{N}} p_l^{\mathbf{a}}(\mathbf{i}, \mathbf{j}) F^*(\mathbf{j}), \ \ \mathbf{i} \in \mathcal{N}. \tag{36}$$

However, these equations are hindered by the *curse of dimensionality*, as the size of the state space $\mathcal{N}$ grows exponentially with the number *L* of projects, which renders them computationally intractable in practice for all but small *L*.

8.2. A Bound for the MAMGBP

Ref. [35] introduced a Lagrangian approach for obtaining a lower bound for the optimal value of the MAMGBP, extending that of Whittle [8] for the case of two-gear projects. First, we construct a *relaxation* of problem (P) by (i) relaxing the class of admissible policies from $\mathbf{\Pi}(\bar{q})$ to $\mathbf{\Pi}(\infty)$, thus allowing violations to the sample-path peak resource constraints

$$q(\mathbf{s}(t), \mathbf{a}(t)) \leqslant \bar{q}, \ \ t = 0, 1, \dots,$$

and (ii) replacing the latter by the following aggregate relaxed version in expectation:

$$\mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{t=0}^{\infty} q(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] \leqslant \frac{\bar{q}}{1 - \beta}.$$

This leads to the following *relaxation* of problem (P) in (35):

$$\begin{aligned} (\widehat{P}) \quad & \text{minimize } \mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{t=0}^{\infty} h(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] \\ & \text{subject to: } \pi \in \mathbf{\Pi}(\infty) \\ & \mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{t=0}^{\infty} q(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] \leqslant \frac{\bar{q}}{1 - \beta}. \end{aligned} \tag{37}$$

The *relaxed problem* $(\widehat{P})$ is a *constrained MDP* (see, e.g., [48]), for which an optimal policy generally depends on the initial state $\mathbf{i}$. Such problems are amenable to a Lagrangian approach. Introducing a non-negative multiplier $\lambda \geqslant 0$ attached to the constraint in (37), we can *dualize* the latter, i.e., bring it into the objective, obtaining the *Lagrangian relaxation*

$$(\widehat{P}_{\lambda}) \quad \underset{\pi \in \mathbf{\Pi}(\infty)}{\text{minimize}} \ \mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{t=0}^{\infty} h(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] + \lambda \left(\mathbb{E}_{\mathbf{i}}^{\pi} \left[\sum_{t=0}^{\infty} q(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] - \frac{\bar{q}}{1 - \beta} \right). \tag{38}$$

Note that, for any initial joint state $\mathbf{i}$ and multiplier $\lambda \geqslant 0$, the optimal cost $\widehat{V}^*(\mathbf{i}, \lambda)$ of $(\widehat{P}_{\lambda})$ is a *lower bound* for that of relaxed problem $(\widehat{P})$, which we denote by $\widehat{F}^*(\mathbf{i})$. In turn, the latter gives a lower bound for the optimal cost $F^*(\mathbf{i})$ of (P). Thus,

$$\widehat{V}^*(\mathbf{i}, \lambda) \leqslant \widehat{F}^*(\mathbf{i}) \leqslant F^*(\mathbf{i}). \tag{39}$$

In light of (39), we are interested in finding an optimal multiplier $\lambda^*(\mathbf{i})$ solving the *dual problem*

$$(D) \quad \underset{\lambda \geqslant 0}{\text{maximize}} \ \widehat{V}^*(\mathbf{i}, \lambda). \tag{40}$$

Since $\widehat{V}^*(\mathbf{i}, \lambda)$ is a *concave function* of λ, being a minimum of linear functions of λ, a local maximum of problem (D) will be a global maximum. Furthermore, since the above problems can be formulated as finite *linear optimization* (LO) problems, the *strong duality* property of the latter ensures the existence of an optimal multiplier $\lambda^*(\mathbf{i}) \geqslant 0$ solving

problem (D) which attains the upper bound $\widehat{F}^*(\mathbf{i})$, i.e., with $\widehat{V}^*(\mathbf{i}, \lambda^*(\mathbf{i})) = \widehat{F}^*(\mathbf{i})$. This corresponds to the satisfaction of the *complementary slackness* property

$$\lambda^*(\mathbf{i}) \left(\mathbb{E}_{\mathbf{i}}^{\pi^*} \left[\sum_{t=0}^{\infty} q(\mathbf{s}(t), \mathbf{a}(t)) \beta^t \right] - \frac{\bar{q}}{1-\beta} \right) = 0, \tag{41}$$

where π^* is an optimal policy for problem $(\widehat{P}_{\lambda^*(\mathbf{i})})$.

Now, since individual project state transitions are conditionally independent given that a joint action has been selected, it suffices —as noted by Whittle [8] for the case of two-gear projects— to consider in $(\widehat{P}_\lambda)$ *decoupled policies* $\boldsymbol{\pi} = (\pi_l)_{l=1}^L$, where $\pi_l \in \Pi_l$ and Π_l is the class of admissible policies for operating project l *as if it were in isolation*. This allows us to reformulate problem $(\widehat{P}_\lambda)$ as

$$(\widehat{P}_\lambda) \quad \text{minimize} \sum_{l=1}^L \mathbb{E}_{i_l}^{\pi_l} \left[\sum_{t=0}^{\infty} \left(h_l(s_l(t), a_l(t)) + \lambda q_l(s_l(t), a_l(t)) \right) \beta^t \right] - \lambda \frac{\bar{q}}{1-\beta} \tag{42}$$

$$\text{subject to: } \pi_l \in \Pi_l, \ \ l = 1, \ldots, L.$$

We can thus *decouple* problem $(\widehat{P}_\lambda)$ into the *individual project subproblems*

$$(\widehat{P}_{l,\lambda}) \quad \underset{\pi_l \in \Pi_l}{\text{minimize}} \ \mathbb{E}_{i_l}^{\pi_l} \left[\sum_{t=0}^{\infty} \left(h_l(s_l(t), a_l(t)) + \lambda q_l(s_l(t), a_l(t)) \right) \beta^t \right], \tag{43}$$

for $l = 1, \ldots, L$. Denoting by $V_l^*(i_l, \lambda)$ the minimum cost objective of subproblem $(\widehat{P}_{l,\lambda})$, it follows that the optimal cost $\widehat{V}^*(\mathbf{i}, \lambda)$ of Lagrangian relaxation $(\widehat{P}_\lambda)$ is decoupled as

$$\widehat{V}^*(\mathbf{i}, \lambda) = \sum_{l=1}^L V_l^*(i_l, \lambda) - \lambda \frac{\bar{q}}{1-\beta}, \tag{44}$$

which allows us to reformulate dual problem (D) in (40) as

$$(D) \quad \underset{\lambda \geqslant 0}{\text{maximize}} \sum_{l=1}^L V_l^*(i_l, \lambda) - \lambda \frac{\bar{q}}{1-\beta} \tag{45}$$

Now, suppose that each project l is indexable with DAI $\lambda_l^*(j_l, a_l)$, so such indices characterize as in Definition 1 the optimal policies for individual project subproblems $(\widehat{P}_{l,\lambda})$, which facilitates the evaluation of optimal costs $V_l^*(i_l, \lambda)$ and hence the computational solution of dual problem (D). For such a purpose, one can use the result that, if $\pi_l^*(\lambda)$ is an optimal policy for subproblem $(\widehat{P}_{l,\lambda})$, then $-\mathbb{E}_{i_l}^{\pi_l^*(\lambda)} \left[\sum_{t=0}^{\infty} q_l(s_l(t), a_l(t)) \beta^t \right]$ is a *subgradient* of $V_l^*(i_l, \lambda)$, seen as a function of λ.

8.3. A Downshift Index Policy for the MAMGBP

Assuming that individual projects are indexable, the author proposed in [35] a suboptimal heuristic index policy for the above MAMGBP based on the projects' DAIs. Here we present a different proposal of a heuristic index policy based on individual project DAIs, which is more easily implementable than that in [35].

Suppose that at time t the joint state is $\mathbf{j} = (j_l)_{l=1}^L$. Consider the project DAIs evaluated at such states, $\lambda_l^*(j_l, a_l)$, for project $l = 1, \ldots, L$. The proposed index policy is described in Algorithm 2, which specifies how to obtain the joint action $\widehat{\mathbf{a}} = (\widehat{a}_l)_{l=1}^L$ prescribed in such a joint state.

Algorithm 2: Downshift index policy for the MAMGBP.

Input: $\mathbf{j} = (j_l)_{l=1}^{L}$ (current joint state)
Output: $\widehat{\mathbf{a}} = (\widehat{a}_l)_{l=1}^{L}$ (prescribed joint action)

Initialization: $a_l := A_l$, $l = 1, \ldots, L$
Loop:
while $\sum_{l=1}^{L} q_l(j_l, a_l) > \bar{q}$ **or** $\min_{1 \leq l \leq L: \, a_l \geq 1} \lambda_l^*(j_l, a_l) \leq 0$ **do**
    pick $\hat{l} \in \arg\min_{1 \leq l \leq L: \, a_l \geq 1} \lambda_l^*(j_l, a_l)$
    $a_{\hat{l}} := a_{\hat{l}} - 1$ (downshift gear in project $\hat{l}$)
end { while }
$\widehat{\mathbf{a}} := \mathbf{a} = (j_l)_{l=1}^{L}$

In short, the algorithm starts by assigning the highest possible gear A_l to each project l. If this is feasible, in that it does not violate the peak resource constraint, this would be the prescribed joint action. Otherwise, the algorithm proceeds by downshifting one of the projects to the next lower gear. The chosen project is one that has minimum DAI at the current gear. The algorithm proceeds until the peak resource constraint is satisfied and the DAIs at the projects with prescribed active actions, if any, are non-negative. In light of the above, we call the policy resulting from this algorithm the *downshift index policy*.

The intuition behind the design of such a policy is that projects should be operated in such a way that two conflicting goals are balanced: (1) higher gears are to be preferred to lower gears whenever possible; and (2) the resulting joint action must be feasible, satisfying the peak resource constraint. The proposed downshift index policy is designed to strike such a balance. If a joint action is not feasible so that a project must be downshifted to a lower gear, the chosen project is one where the loss in performance due to such a downshift, for which the project DAIs are used as a proxy measure, is minimal.

Note that the downshift index policy reduces to the Whittle index policy in the case of two-gear projects.

9. Some Extensions

This section presents some extensions to the above framework.

9.1. Extension to the Long-Run Average Cost Criterion

The above results for the discounted cost criterion readily extend to the *(long-run) average cost criterion* (see, e.g., ([47] Ch. 8)) under appropriate *ergodicity* conditions. Consider the average cost, including holding and resource usage costs (charged at price λ), of running the project starting from state i under a policy $\pi \in \Pi$, defined by

$$\overline{V}_i(\lambda, \pi) \triangleq \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_i^{\pi} \left[\sum_{t=0}^{T-1} \left(h_{s(t)}^{a(t)} + \lambda q_{s(t)}^{a(t)} \right) \right].$$

We further define the corresponding optimal cost

$$\overline{V}_i^*(\lambda) \triangleq \inf_{\pi \in \Pi} \overline{V}_i(\lambda, \pi).$$

We can thus formulate the project's *average λ-price problem* as

$$(\overline{P}_\lambda) \quad \text{find } \pi^*(\lambda) \in \Pi: \overline{V}_i(\lambda, \pi^*(\lambda)) = \overline{V}_i^*(\lambda), \quad i \in \mathcal{N}. \tag{46}$$

We shall refer to a policy $\pi^*(\lambda)$ solving the average λ-price problem $(\overline{P}_\lambda)$ as a *λ-optimal policy*.

We shall make the following assumption.

Assumption 2. *The following conditions hold:*

(i) *The model is weakly accessible, so the state space $\mathcal{N}$ can be partitioned into two subsets $\mathcal{N}^{\mathrm{tr}}$ and $\mathcal{N}^{\mathrm{acc}}$, such that (i.a) all states in $\mathcal{N}^{\mathrm{tr}}$ are transient under every stationary policy and (i.b) for every two states i and j in $\mathcal{N}^{\mathrm{acc}}$, j is accessible from i, so there exists a stationary policy π and a positive integer t such that $\mathbb{P}_i^\pi\{s(t) = j\} > 0$.*

(ii) *Every policy $S \in \mathcal{F}$ is unichain, i.e., it induces a single recurrent class plus possible additional transient states.*

Now, by standard results in average-cost MDP theory (see [49] (Sec. 5.2)), Assumption 2 (i) ensures the existence of a λ-optimal policy $\pi^*(\lambda) \in \Pi^{\mathrm{SD}}$, with the optimal average cost $\overline{V}_i^*(\lambda)$ being independent of the initial state, i.e., $\overline{V}_i^*(\lambda) \equiv \overline{V}^*(\lambda)$.

Now, by using the *Laurent series expansions* for finite-state and -action MDP models (see Corollaries 8.2.4 and 8.2.5 in [47]), we have the following. For any stationary deterministic policy $\pi \in \Pi^{\mathrm{SD}}$,

$$\overline{x}_{ij}^a(\pi) \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_i^\pi\left[\sum_{t=0}^{T-1} 1_{\{a(t)=a\}}\right] = \lim_{\beta\nearrow 1}(1-\beta)x_{ij}^a(\pi),$$

$$\overline{F}_i(\pi) \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_i^\pi\left[\sum_{t=0}^{T-1} h_{s(t)}^{a(t)}\right] = \lim_{\beta\nearrow 1}(1-\beta)F_i(\pi),$$

$$\overline{G}_i(\pi) \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_i^\pi\left[\sum_{t=0}^{T-1} q_{s(t)}^{a(t)}\right] = \lim_{\beta\nearrow 1}(1-\beta)G_i(\pi).$$

Furthermore, for any $S = (S_0, \ldots, S_A) \in \mathcal{F}$, Assumption 2(ii) ensures that the above metrics do not depend on the initial state i, so we can write $\overline{x}_j^a(S)$, $\overline{F}(S)$, and $\overline{G}(S)$. Furthermore, we have the Laurent series expansions

$$F_i(S) = \frac{\overline{F}(S)}{1-\beta} + \varphi_i(S) + O(1-\beta), \ \text{ as } \ \beta \nearrow 1$$

and

$$G_i(S) = \frac{\overline{G}(S)}{1-\beta} + \gamma_i(S) + O(1-\beta), \ \text{ as } \ \beta \nearrow 1,$$

where the *bias* terms $\varphi_i(S)$ and $\gamma_i(S)$ are determined, up to an additive constant, by the evaluation equations

$$\overline{F}(S) + \varphi_i(S) = h_i^a + \sum_{j\in\mathcal{N}} p_{ij}^a \varphi_j(S), \ \ i \in S_a, a \in \mathcal{A}$$

and

$$\overline{G}(S) + \gamma_i(S) = q_i^a + \sum_{j\in\mathcal{N}} p_{ij}^a \gamma_j(S), \ \ i \in S_a, a \in \mathcal{A}.$$

From the above, (9) and (10) we can define the average *marginal (holding) cost metric*

$$\overline{f}_i^{a,a'}(S) \triangleq h_i^a - h_i^{a'} + \sum_{j\in\mathcal{N}} p_{ij}^a \varphi_j(S) - \sum_{j\in\mathcal{N}} p_{ij}^{a'} \varphi_j(S) = \lim_{\beta\nearrow 1} f_i^{a,a'}(S) \tag{47}$$

and the the average *marginal resource (usage) metric*

$$\overline{g}_i^{a,a'}(S) \triangleq q_i^{a'} - q_i^a + \sum_{j\in\mathcal{N}} p_{ij}^{a'} \gamma_j(S) - \sum_{j\in\mathcal{N}} p_{ij}^a \gamma_j(S) = \lim_{\beta\nearrow 1} g_i^{a,a'}(S). \tag{48}$$

If $\bar{g}_i^{a,a'}(S) > 0$, we further define the *average MP metric*

$$\overline{m}_i^{a,a'}(S) \triangleq \frac{\bar{f}_i^{a,a'}(S)}{\bar{g}_i^{a,a'}(S)} = \lim_{\beta \nearrow 1} m_i^{a,a'}(S). \tag{49}$$

We thus have the following verification theorem, which is the average criterion counterpart to Theorem 1 for the discounted criterion. Note that the following theorem refers to the corresponding concepts for the average criterion of $\mathcal{F}$-indexability, PCL($\mathcal{F}$)-indexability, and downshifting algorithm $\overline{\mathrm{DS}}(\mathcal{F})$, which is as algorithm DS($\mathcal{F}$) but using the average marginal metrics instead of the discounted ones.

Theorem 2. *If the average cost model is PCL($\mathcal{F}$)-indexable, then it is $\mathcal{F}$-indexable, with DAI $\bar{\lambda}_j^{*,a}$ given by the MPI $\overline{m}_j^{*,a}$.*

9.2. Models with Uncontrollable States

In the above framework, we have assumed that the DAI $\lambda_j^{*,a}$ is defined for all project states $j \in \mathcal{N}$. Yet, in some models, this need not be the case, in particular in those having *uncontrollable states*. We call a project state i *uncontrollable* if only one action is available at i, or, equivalently, if all actions a give the same transition probabilities, so $p_{ij}^a = p_{ij}^0$ for all a. This concept was considered by the author in the corresponding framework for two-gear projects developed in [29]. If there are uncontrollable states, we decompose the state space as $\mathcal{N} = \mathcal{N}^{\mathrm{cont}} \cup \mathcal{N}^{\mathrm{unc}}$, where $\mathcal{N}^{\mathrm{cont}}$ is the controllable state space and $\mathcal{N}^{\mathrm{unc}}$ is the uncontrollable state space.

In such a case, the above framework carries over by defining the concept of indexability and DAI by focusing on the controllable state space $\mathcal{N}^{\mathrm{cont}}$, so the DAI $\lambda_j^{*,a}$ will only be defined for states $j \in \mathcal{N}^{\mathrm{cont}}$. The required adaptions are straightforward. For example, the policy notation $S = (S_0, \dots, S_A)$ used above can now be interpreted as meaning that, under such a policy, action a is taken in controllable states $j \in S_a$ for $a = 0, \dots, A$, as $S_0, \dots, S_A$ is now a partition of $\mathcal{N}^{\mathrm{cont}}$.

9.3. Models with a Countably Infinite State Space

The extension of the above framework to models with a countably infinite state space raises issues mainly in the definition of the downshift adaptive-greedy algorithm. Thus, the algorithm would not terminate, and it might not traverse the entire space of (j, a) for which the index is defined. Furthermore, it might possibly entail choosing among infinitely many state–action pairs (j, a) at each step.

Yet, in some countably infinite state space models such issues are easily addressed. Consider, e.g., a model that might arise in queueing theory where the state is the number of customers in the system so the state space is the set of non-negative integers, $\mathcal{N} \triangleq \{0, 1, 2, \dots\}$. Imagine that the actions or gears a correspond to server speeds, so higher gears give faster service rates. The holding costs are used to model penalties (possibly nonlinear) for congestion.

In such a setting, it is natural to conjecture that optimal policies should be *multi-threshold policies*. Any such policy is characterized by thresholds $z^1 \leqslant z^2 \leqslant \cdots \leqslant z^A$, with the interpretation that gear 0 is used in states $1 \leqslant j \leqslant z^1$, gear a is used in states $z^a < j \leqslant z^{a+1}$ for $a = 1, \dots, A-1$, and gear A is used in states $j > z^A$. Note that the optimality of such policies has been established in some queueing models, see, e.g., [40–43].

The present framework would be applied to such a setting as follows. Rather than considering directly such multi-threshold policies and trying to establish their optimality, one would postulate a corresponding family of policies $\mathcal{F}$. Note that in such a model state 0 would be uncontrollable, as there is no meaningful choice of action when the queue is empty. Thus, excluding state 0 from consideration, the postulated family $\mathcal{F}$ would consist of partitions $S = (S_0, \dots, S_A)$ of the controllable state space with the following property: if

gear a is selected in a state j (i.e., $j \in S_a$), then at any lower state $j' < j$ a gear $a' \leqslant a$ must be selected.

It is easy to see that, in this setting, the natural extension of the downshift adaptive-greedy algorithm would indeed traverse the entire space of state–action pairs (j, a) for which the DAI is defined, which would provide an alternative approach to address such problems to that previously considered in the aforementioned literature.

10. Discussion

This paper has introduced novel sufficient conditions for the indexability of multi-gear bandits modeling a dynamic and stochastic project consuming a single resource, along with an efficient index-computing algorithm. This can be used to efficiently solve general MDP models that satisfy such conditions, and the index has further been used to design a heuristic index policy for the more complex multi-armed multi-gear bandit problem. This work opens a number of further avenues for developing such an approach, including the following: developing an efficient implementation of and testing the algorithm; deploying the new PCL-indexability conditions in a variety of relevant models arising in applications; extending the approach to models with a countable state space; and extending the approach to models with a continuous state space.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bellman, R. A problem in the sequential design of experiments. *Indian J. Stat.* **1956**, *16*, 221–229.
2. Gittins, J.C.; Jones, D.M. A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics (European Meeting of Statisticians, Budapest, 1972)*; Gani, J., Sarkadi, K., Vincze, I., Eds.; North-Holland: Amsterdam, The Netherlands, 1974; pp. 241–266.
3. Gittins, J.C. Bandit processes and dynamic allocation indices. *J. Roy. Stat. Soc. Ser. B* **1979**, *41*, 148–177. [CrossRef]
4. Gittins, J.C. *Multi-Armed Bandit Allocation Indices*; Wiley: Chichester, UK, 1989.
5. Whittle, P. Multi-armed bandits and the Gittins index. *J. Roy. Stat. Soc. Ser. B* **1980**, *42*, 143–149. [CrossRef]
6. Weber, R. On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.* **1992**, *2*, 1024–1033. [CrossRef]
7. Bertsimas, D.; Niño-Mora, J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.* **1996**, *21*, 257–306. [CrossRef]
8. Whittle, P. Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.* **1988**, *25*, 287–298. [CrossRef]
9. Veatch, M.H.; Wein, L.M. Scheduling a multiclass make-to-stock queue: Index policies and hedging points. *Oper. Res.* **1996**, *44*, 634–647. [CrossRef]
10. Niño-Mora, J. Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue. *Queueing Syst.* **2006**, *54*, 281–312. [CrossRef]
11. Niño-Mora, J. Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging. In Proceedings of the International Conference on Network Control and Optimization (NET-COOP 2007), Avignon, France, 5–7 June 2007; Chahed, T., Tuffin, B., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4465, pp. 138–149.
12. Kumar, U.D.; Saranga, H. Optimal selection of obsolescence mitigation strategies using a restless bandit model. *Eur. J. Oper. Res.* **2010**, *200*, 170–180. [CrossRef]
13. Washburn, R.B.; Schneider, M.K. Optimal policies for a class of restless multiarmed bandit scheduling problems with applications to sensor management. *J. Adv. Inform. Fusion* **2008**, *3*, 3–13.
14. Liu, K.; Zhao, Q. Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access. *IEEE Trans. Inform. Theory* **2010**, *56*, 5547–5567. [CrossRef]
15. Borkar, V.S.; Kasbekar, G.S.; Pattathil, S.; Shetty, P.Y. Opportunistic scheduling as restless bandits. *IEEE Trans. Control Netw. Syst.* **2018**, *5*, 1952–1961. [CrossRef]

16. Yang, F.; Luo, X. A restless MAB-based index policy for UL pilot allocation in massive MIMO over Gauss–Markov fading channels. *IEEE Trans. Veh. Technol.* **2020**, *69*, 3034–3047. [CrossRef]

17. Abbou, A.; Makis, V. Group maintenance: A restless bandits approach. *Informs J. Comput.* **2019**, *31*, 719–731. [CrossRef]

18. La Scala, B.F.; Moran, B. Optimal target tracking with restless bandits. *Digit. Signal Process.* **2006**, *16*, 479–487. [CrossRef]

19. Dance, C.R.; Silander, T. Optimal policies for observing time series and related restless bandit problems. *J. Mach. Learn. Res.* **2019**, *20*, 35.

20. Niño-Mora, J. A faster index algorithm and a computational study for bandits with switching costs. *Informs J. Comput.* **2008**, *20*, 255–269. [CrossRef]

21. Niño-Mora, J. Fast two-stage computation of an index policy for multi-armed bandits with setup delays. *Mathematics* **2021**, *9*, 52. [CrossRef]

22. Ayer, T.; Zhang, C.; Bonifonte, A.; Spaulding, A.C.; Chhatwal, J. Prioritizing hepatitis C treatment in US prisons. *Oper. Res.* **2019**, *67*, 853–873. [CrossRef]

23. Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; Tambe, M. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. *arXiv* **2019**, arXiv:2109.08075.

24. Fu, J.; Moran, B.; Taylor, P.G. A restless bandit model for resource allocation, competition, and reservation. *Oper. Res.* **2022**, *70*, 416–431. [CrossRef]

25. Niño-Mora, J. Restless bandits, partial conservation laws and indexability. *Adv. Appl. Probab.* **2001**, *33*, 76–98. [CrossRef]

26. Klimov, G.P. Time-sharing service systems. I. *Theory Probab. Appl.* **1974**, *19*, 532–551. [CrossRef]

27. Coffman, E.G., Jr.; Mitrani, I. A characterization of waiting time performance realizable by single-server queues. *Oper. Res.* **1980**, *28*, 810–821. [CrossRef]

28. Shanthikumar, J.G.; Yao, D.D. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Oper. Res.* **1992**, *40*, S293–S299. [CrossRef]

29. Niño-Mora, J. Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Math. Program.* **2002**, *93*, 361–413. [CrossRef]

30. Niño-Mora, J. Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock M/G/1 queues. *Math. Oper. Res.* **2006**, *31*, 50–84. [CrossRef]

31. Niño-Mora, J. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP* **2007**, *15*, 161–198. [CrossRef]

32. Niño-Mora, J. A verification theorem for threshold-indexability of real-state discounted restless bandits. *Math. Oper. Res.* **2020**, *45*, 465–496. [CrossRef]

33. Niño-Mora, J. A fast-pivoting algorithm for Whittle's restless bandit index. *Mathematics* **2020**, *8*, 2226. [CrossRef]

34. Weber, R. Comments on: Dynamic priority allocation via restless bandit marginal productivity indices. *TOP* **2007**, *15*, 211–216. [CrossRef]

35. Niño-Mora, J. An index policy for multiarmed multimode restless bandits. In Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools (ValueTools'08), Athens, Greece, 20–24 October 2008; ICST: Brussels, Belgium, 2008. [CrossRef]

36. Zayas-Cabán, G.; Jasin, S.; Wang, G. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Adv. Appl. Probab.* **2019**, *51*, 745–772. [CrossRef]

37. Killian, J.A.; Biswas, A.; Shah, S.; Tambe, M. Q-learning Lagrange policies for multi-action restless bandits. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; ACM: New York, NY, USA, 2021; pp. 871–881.

38. Serfozo, R.F. Monotone optimal policies for Markov decision processes. In *Stochastic Systems: Modeling, Identification and Optimization, II*; Wets, R.J.B., Ed.; Mathematical Programming Studies; Springer: Berlin/Heidelberg, Germany, 1976; Volume 6, pp. 202–215.

39. Heyman, D.P.; Sobel, M.J. *Stochastic Models in Operations Research. Vol. II: Stochastic Optimization*; McGraw-Hill: New York, NY, USA, 1984.

40. Crabill, T.B. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Manag. Sci.* **1972**, *18*, 560–566. [CrossRef]

41. Sabeti, H. Optimal selection of service rates in queueing with different cost. *J. Oper. Res. Soc. Jpn.* **1973**, *16*, 15–35.

42. Ata, B.; Shneorson, S. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Manag. Sci.* **2006**, *52*, 1778–1791. [CrossRef]

43. Mayorga, M.E.; Ahn, H.S.; Shanthikumar, J.G. Optimal control of a make-to-stock system with adjustable service rate. *Probab. Eng. Inform. Sci.* **2006**, *20*, 609–634. [CrossRef]

44. Ye, Y.Y. The simplex and policy-iteration methods are strongly polynomial for the Markov Decision Problem with a fixed discount rate. *Math. Oper. Res.* **2011**, *36*, 593–603. [CrossRef]

45. Scherrer, B. Improved and generalized upper bounds on the complexity of policy iteration. *Math. Oper. Res.* **2016**, *41*, 758–774. [CrossRef]

46. Hollanders, R.; Gerencser, B.; Delvenne, J.C.; Jungers, R.M. Improved bound on the worst case complexity of Policy Iteration. *Oper. Res. Lett.* **2016**, *44*, 267–272. [CrossRef]

47. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; Wiley: New York, NY, USA, 1994.
48. Altman, E. *Constrained Markov Decision Processes*; Chapman & Hall/CRC: Boca Raton, FL, USA, 1999.
49. Bertsekas, D.P. *Dynamic Programming and Optimal Control*, 4th ed.; Athena Scientific: Belmont, MA, USA, 2012; Volume 2.