The LP-update policy for weakly coupled MDPs*

Chen Yan[1], Nicolas Gast[1], Bruno Gaujal[1]

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG

{chen.yan,nicolas.gast,bruno.gaujal}@inria.fr

Résumé / Abstract : *Weakly coupled MDPs generalize the notion of resltess multi-armed bandits, that are known to be PSPACE-hard. In this paper, we show that a relaxed version of the problem can be solved by linear programming. We use the solution of a relaxed problem to build a heuristics, that we call the LP-update policy. We study its performance for a case with N statistically identical weakly coupled MDPs. We prove that, when N goes to infinity, the LP-update policy is asymptotically optimal. We also provide a rank condition that guarantees that the sub-optimality gap is at most $O(1/N)$. We illustrate the good performance of our algorithm in a selection problem with fairness constraints.*

Mots-clés / Keywords : *Markov decision processes, stochastic optimization, linear programming, restless bandit*

1 Introduction

Markov decision processes (MDPs) have proven tremendously useful as models of stochastic sequential planning problems. Dynamic programming is the classical method to solve MDPs, but requires the MDP to have a small state and action space, which is often not the case. Hence, the computational difficulty of applying classic dynamic programming algorithms to realistic problems has triggered much research into techniques to deal with large state and action spaces. One such technique is *decomposition*, for which the very large global MDP is decomposed into N loosely dependent sub-processes. Each of these N sub-problems is exponentially smaller in size compared to the global problem. If these solutions can be pieced together effectively, and used to guide the search for a global solution that performs well, then dramatic improvements in the overall solution time can be obtained.

Weakly coupled MDPs fall into this situation. The model originates from sequential stochastic resource allocation problems: A number of tasks must be addressed and actions consists of assigning various resource at every decision epoch to each of these tasks. The tasks are *additive utility independent*: the utility of any collection of tasks is the sum of rewards associated with each task. In addition, each task can be viewed as an independent sub-process whose rewards and transitions are independent of the others, given a fixed action or policy. These tasks are only coupled via the global resource constraints at each decision epoch. This explains the terminology "weakly coupled". Weakly coupled MDPs are widely applicable in practice, and can be used to model queueing [1], scheduling [5] or health care [6] problems.

In this work we propose a novel policy called the *LP-update policy* for finite horizon weakly coupled MDPs. We consider a scaling model with N statistically identical components – each *component* representing for instance a task to complete – which are coupled by a set of constraints that must hold at all time. We show that the relaxed problem in which the constraints must only be satisfied in expectation can be solved by an LP, and we demonstrate how to use the solution of this LP to build the LP-update policy. Our main result is to show that the LP update policy is asymptotically optimal as N goes to infinity. We show that the sub-optimality gap is $O(1/\sqrt{N})$ for all problems, and might be reduced to $O(1/N)$ for problems

[*] A long version of this work containing a full proof of the theorem as well as additional insights on weakly coupled MDPs is available in [4].

that are *non-degenerate*. A problem is said to be non-degenerate if the saturated constraints of the LP satisfies some rank conditions. This notion was already defined for classical restless bandits [3, 9] and one of our contributions has been to extend it for any weakly coupled MDP. We conclude our paper by studying a candidate selection problem with fairness constraints. We compare the performance of the LP-update policy and the occupation measure policy of [8], that is known to be asymptotically optimal, but with a sub-optimality gap of $O(1/\sqrt{N})$ for all problems. Our results show that for finite N, the LP-update clearly outperforms the occupation measure policy, and it better exploits the structural properties of the problems.

2 Model description

We consider a finite-horizon discrete-time weakly coupled MDP composed of N statistically identical sub-MDPs (or *components* in the bandit literature), indexed by $n \in \{1 \dots N\}$. The finite state space of each sub-MDP is the set $\mathcal{S} := \{1, 2, \dots, d\}$, and its finite action space is $\mathcal{A} := \{0, 1, \dots, A\}$. The state space of the weakly coupled MDP is therefore $\mathcal{S}^N$ and the action space is a subset of $\mathcal{A}^N$. There are J types of resources, and the decision maker is allowed to use up to b_j resources of type j at each decision epoch. We assume that taking the action a_n for component n that is in state s_n uses $D_j(s_n, a_n) \geq 0$ of resource j, and that the action 0 consumes no resource: $D_j(s_n, 0) = 0$ for all $s_n \in \mathcal{S}$. Hence, the set of *feasible* actions in state $\mathbf{s} = (s_1, s_2, \dots, s_N)$ is the set of $\mathbf{a} \in \mathcal{A}^N$ such that for all $j \in \{1 \dots J\}$: $\sum_{n=1}^{N} D_j(s_n, a_n) \leq N \cdot b_j$.

Upon choosing an action $\mathbf{a}$ in state $\mathbf{s}$, the decision maker receives a reward $\sum_{n=1}^{N} R_{s_n}^{a_n}$. We assume that the sub-processes are weakly coupled, in the sense that the N sub-MDPs are only linked through the J resource constraints. For a given feasible action $\mathbf{a}$, the system transitions from a state $\mathbf{s}$ to state $\mathbf{s}' = (s'_1, s'_2, \dots, s'_N)$ with probability

$$p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \prod_{n=1}^{N} p(s'_n \mid s_n, a_n) = \prod_{n=1}^{N} P_{s_n, s'_n}^{a_n}, \tag{1}$$

where for each action a, the matrix $\mathbf{P}^a$ is a probability transition matrix of dimension $d \times d$.

Remark that this model includes the classical restless multi-armed bandit that corresponds to the case $\mathcal{A} = \{0, 1\}$ with $J = 1$ constraint and $D_1(s, a) = a$. This simpler optimization problem is already PSPACE-hard (see [7]). We hence focus on developing approximate solutions whose performance are provably close to optimal.

3 The LP relaxation

As all sub-MDPs are statistically indentical, we denote by $M_s^{(N)}(t)$ the proportion of sub-MDPs that are in state $s \in \mathcal{S}$ at decision epoch t. $\mathbf{M}^{(N)}(t) = (M_s^{(N)}(t))_{s \in \mathcal{S}}$ is the state vector, and $\Delta^{(N),d}$ is the set of possible values for $\mathbf{M}^{(N)}$, which is the simplex of dimension d where each coordinate is a multiple of $1/N$. Let $\mathcal{U} = \mathcal{S} \times \mathcal{A}$ be the set of state-action pairs and let $u = |\mathcal{U}|$. A possible action is a vector $\mathbf{y} \in \mathbb{R}^u$ where $y_{s,a}$ is the proportion of sub-MDPs that are in state s and for which action a is taken. We say that $\mathbf{Y}^{(N)}$ is admissible for $\mathbf{M}^{(N)}$ if $\mathbf{Y}^{(N)} \geq 0$, $D\mathbf{Y}^{(N)} \leq d$, $\sum_a Y_{s,a}^{(N)} = M_s^{(N)}$ and $NY_{s,a}^{(N)}$ is an integer for all s, a.

The decision maker's goal is to maximize the total expected reward over a finite horizon T, given an initial configuration $\mathbf{m}(0) \in \Delta^{(N),d}$ of the system. Its value $V_{\text{opt}}^{(N)}(\mathbf{m}(0), T)$ equals

$$V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) = \max_{\mathbf{Y}} \quad \mathbb{E}\Big[\sum_{t=0}^{T-1} \sum_{(s,a) \in \mathcal{U}} R_s^a Y_{s,a}^{(N)}(t)\Big] \tag{2a}$$

$$\text{s.t.} \quad \mathbf{M}^{(N)}(0) = \mathbf{m}(0), \tag{2b}$$

$$\mathbf{Y}^{(N)}(t) \text{ is admissible for } \mathbf{M}^{(N)}(t) (\forall t), \tag{2c}$$

$$\mathbf{M}^{(N)}(t+1) \text{ follows the Markov transitions (1) given } \mathbf{Y}^{(N)}(t). \tag{2d}$$

The main difficulty of the above optimization problem is that the constraint $D\mathbf{Y}^{(N)}(t) \leq \mathbf{b}$ couples all sub-MDPs. This makes the problem computationally hard. To overcome this difficulty, we use a method similar to the one of [2, 3, 9] and relax these constraints by assuming that they should be satisfied only in expectation: $D\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] \leq \mathbf{b}$. This leads us to write a relaxed optimization problem in terms of the variables $\mathbf{y}(t) = \mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right]$ because (1) implies that the expectation of $\mathbf{M}^{(N)}(t+1)$ given $\mathbf{Y}^{(N)}(t)$ can be rewritten as a linear map:

$$\mathbb{E}\left[M_s^{(N)}(t+1) \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] = \sum_{(s',a)\in\mathcal{U}} y_{s',a}P^a_{s',s} := (\mathbf{y}\mathbf{P})_s. \tag{3}$$

The value of this relaxed problem, $V_{\text{rel}}(\mathbf{m}(0), T)$, is the solution of the following LP:

$$\begin{aligned} V_{\text{rel}}(\mathbf{m}(0), T) = \ \max \quad & \sum_{t=0}^{T-1} \sum_{(s,a)\in\mathcal{U}} R^a_s y_{s,a}(t) && && \text{(4a)} \\ \text{s.t.} \quad & \sum_{a\in\mathcal{A}} y_{s,a}(0) = m_s(0) && \forall s, && \text{(4b)} \\ & D\mathbf{y}(t) \leq \mathbf{b} \text{ and } \mathbf{y}(t) \geq \mathbf{0} && \forall t, && \text{(4c)} \\ & \sum_{a\in\mathcal{A}} y_{s,a}(t+1) = (\mathbf{y}(t)\mathbf{P})_s && \forall s, t. && \text{(4d)} \end{aligned}$$

In the above formulation, the constraint (4b) corresponds to the condition on the initial state (2b), the constraint (4c) is the relaxed version of (2c), and (4d) comes from (2d), (3), and the fact that $M_s(t+1) = \sum_a Y_{s,a}(t+1)$.

By the assumptions that $D(s, 0) = \mathbf{0}$ and $D(s, a), \mathbf{b} \geq \mathbf{0}$, the linear program (4) is feasible (*e.g.* it suffices to always choose the passive action $a = 0$). In the following, we denote by $\mathbf{y}^*$ one of its optimal solution, and by $\mathbf{m}^*$ the sequence of vectors $\mathbf{m}^*(t)$ such that $m^*_s(t) = \sum_{a\in\mathcal{A}} y^*_{s,a}(t)$.

4 LP-update policies, non-degeneracy and suboptimality gap

4.1 The LP-“full-update” policy

As $V_{\text{rel}}(\mathbf{m}(0), T)$ is a relaxed version of the original problem (2), it should be clear that $V^{(N)}_{\text{opt}}(\mathbf{m}(0), T) \leq V_{\text{rel}}(\mathbf{m}(0), T)$. In what follows, we explain how to construct a policy for the finite N system whose performance converges to $V_{\text{rel}}(\mathbf{m}(0), T)$ as N goes to infinity. In particular, this implies that this policy is asymptotically optimal.

At time t, the decision maker observes $\mathbf{M}^{(N)}(t)$ and must choose an admissible action $\mathbf{Y}^{(N)}(t)$. The optimal decision for the LP is to take $\mathbf{y}^*(t)$. By construction, this optimal solution is feasible for $\mathbf{m}^*(t)$ but is in general not feasible for $\mathbf{M}^{(N)}(t)$ when $t \geq 1$ (mainly because $\mathbf{M}^{(N)}(t)$ is a random variable). A possible approach (which is very similar to the ones developed in [2, 3, 8, 9]) is to set $\mathbf{Y}^{(N)}(t)$ as the admissible action that is the closest to $\mathbf{y}^*(t)$. Yet, this approach neglects the fact that $\mathbf{M}^{(N)}(t)$ might be very far from $\mathbf{m}^*(t)$ because of random fluctuations.

Our approach in this paper: since we know that *at time* 0, $\mathbf{M}^{(N)}(0) = \mathbf{m}(0)$, we know, up to an integer constraint, that $\mathbf{y}^*(0)$ is feasible at time 0. Hence, our algorithm, that we call the *LP-update* policy, solves a new LP starting from $\mathbf{M}^{(N)}(t)$ with horizon $T - t$. This guarantees that the newly computed $\mathbf{y}^*(t)$ is admissible for $\mathbf{M}^{(N)}(t)$, by constraint (4b). Since $Ny^*_{s,a}(t)$ is not necessarily an integer, the idea is then to use a rounding procedure. A simple way to do so is to use $Y^{(N)}_{s,a}(t) := N^{-1}\left\lfloor Ny^*_{s,a}(t)\right\rfloor$. This algorithm is called LP-“full-update” because it updates the solution of the LP at every time-iteration.

4.2 Non-degenerate problems

We decompose the LP (4) into a T optimization problems:

$$
\begin{aligned}
V_{\mathrm{rel}}(\mathbf{m}(t), T-t) = \max_{\mathbf{y} \in \mathbb{R}^u} \quad & \mathbf{R}^\top \mathbf{y} + V_{\mathrm{rel}}(\mathbf{y}\mathbf{P}, T-t-1) && \text{(5a)} \\
\text{s.t.} \quad & \mathbf{y} \geq \mathbf{0}, && \text{(5b)} \\
& D\mathbf{y} \leq \mathbf{b}, && \text{(5c)} \\
& E\mathbf{y} = \mathbf{m}(t), && \text{(5d)}
\end{aligned}
$$

where E is the matrix encoding the equality constraints $\sum_{a \in \mathcal{A}} y_{s,a} = m_s(t)$.

The notion of non-degenerate problems comes by looking at the saturated constraints of (5a). Let $\mathbf{y}^*$ be an optimal solution to the linear program (4) and define the sets:

- $\mathcal{U}^*(t) = \{(s,a) \text{ such that } y^*_{s,a}(t) = 0\}$ (saturated constraints of (5b)).
- $\mathcal{J}^*(t) = \{j : (D\mathbf{y}^*(t))_j = 0\}$ (saturated constraints of (5c)).
- $\mathcal{S}^*(t) = \{s : m^*_s(t) \neq 0\}$ (constraint of (5d) not already saturated in (5b).)

The three sets above are equalities. Hence, they can be represented as a linear constraint

$$
C^*(t)\mathbf{y} = \left[\mathbf{0}|_{\mathcal{U}^*(t)};\ \mathbf{b}|_{\mathcal{J}^*(t)};\ \mathbf{m}(t)|_{\mathcal{S}^*(t)}\right]^\top, \tag{6}
$$

where $C^*(t)$ is a matrix with $|\mathcal{J}^*(t)| + |\mathcal{S}^*(t)| + |\mathcal{U}^*(t)|$ rows and $u = |\mathcal{U}|$ columns. The notations $\mathbf{b}|_{\mathcal{J}^*(t)}$ (or $\mathbf{0}|_{\mathcal{U}^*(t)}$ and $\mathbf{m}(t)|_{\mathcal{S}^*(t)}$) indicate that the vector is restricted to the indices $\mathcal{J}^*(t)$.

It should be clear that adding the constraints (6) to the optimization problem (5) will not change its value nor its optimal solution $\mathbf{y}^*(t)$ because these constraints were already saturated for $\mathbf{y}^*(t)$. However, when $\mathbf{m}$ deviates from $\mathbf{m}^*(t)$, forcing the equalities of (6) might not be optimal for (5), because the saturated constraints of (5) might vary.

In our paper [4], we say that a problem is **non-degenerate**, if the matrix $C^*(t)$ admits a right-inverse for all t (which is to say that $C^*(t)$ has rank $|\mathcal{J}^*(t)| + |\mathcal{S}^*(t)| + |\mathcal{U}^*(t)|$). This shows that if a problem is non-degenerate, the saturated constraints remains identical in a neighborhood of $\mathbf{m}^*(t)$. This implies that the optimal solution of (5) is locally linear around $\mathbf{m}^*(t)$, *i.e.*, there exists a matrix $C^+(t)$ such that for all $\mathbf{m}$ in a neighborhood of $\mathbf{m}^*(t)$, the quantity $\mathbf{y}^*(t) + C^+(t)(\mathbf{m} - \mathbf{m}^*(t))$ is a solution of (5) (the matrix $C^+(t)$ is essentially the right-inverse of $C^*(t)$ restricted to $\mathcal{U}^*(t)$, see [4] for a more precise definition).

Algorithm 1: The improved LP-update policy for weakly coupled MDPs.

```
Input: Initial configuration vector $\mathbf{M}^{(N)}(0) = \mathbf{m}(0)$ over time span $[0, T]$.
for $t = 0, 1, 2, \dots, T-1$ do
    if $t > 0$ and $C^*(t)$ has a right inverse $C^+(t)$ and the quantity
     $\mathbf{y}^*(t) + C^+(t)(\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t))$ is admissible for $\mathbf{M}^{(N)}(t)$ then
        Set $\mathbf{y}(t) := \mathbf{y}^*(t) + C^+(t)(\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t))$.
    else
        Solve LP (4) with initial configuration vector $\mathbf{M}^{(N)}(t)$ over time span $[t, T]$.;
        Set $\mathbf{y}(t) := \mathbf{y}^*(t)$;
    end
    Set $Y^{(N)}_{s,a}(t) := N^{-1} \lfloor N y_{s,a}(t) \rfloor$ for $a \neq 0$ and $Y^{(N)}_{s,0}(t) := M^{(N)}_s(t) - \sum_{a \neq 0} Y^{(N)}_{s,a}(t)$ ;
    Use actions $Y^{(N)}_{s,a}(t)$ over all sub-MDPs to advance to the next time-step;
end
```

This shows that instead of solving a new LP at time t, one can try to compute the right-inverse $C^+(t)$ of $C^*(t)$ and apply the decision $\mathbf{Y}^{(N)}(t) = \mathbf{y}^*(t) + C^+(t)(\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t))$. If $C^*(t)$ does not have a right-inverse or if $\mathbf{y}^*(t) + C^+(t)(\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t))$ is not admissible for $\mathbf{M}^{(N)}(t)$, then we need to solve a new LP. This leads to our improved LP-update policy detailed in Algorithm 1.

4.3 Characterization of the sub-optimality gap

Our main theoretical result is to show that the LP-update is asymptotically optimal for all weakly coupled MDPs. Moreover, the rate at which the LP-update becomes optimal is faster for non-degenerate problems.

Theorem 1 *Denote by $V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0), T)$ the value of the LP-update policy defined in Algorithm 1, and by $V_{\mathrm{rel}}(\mathbf{m}(0), T)$ the value of the linear program (4). For any weakly coupled MDP with statistically identical components, there exists a constant $C > 0$ such that for all N:*

$$\left|V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0), T) - V_{\mathrm{rel}}(\mathbf{m}(0), T)\right| \leq \frac{C}{N^{\alpha}},$$

where $\alpha = 1$ if the problem is non-degenerate, and $\alpha = 0.5$ otherwise.

The proof of this result is given in [4], where we also define a notion of *perfect rounding* for which the suboptimality gap is exponentially small (of order $e^{-\Omega(N)}$). For all three cases, we show that these bounds are tight, by giving examples of problems such that $\left|V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0), T) - V_{\mathrm{rel}}(\mathbf{m}(0), T)\right|$ is at least $\Omega(1/\sqrt{N})$, at least $\Omega(1/N)$ or at least $e^{-O(N)}$.

5 Case study: Impact of fairness in selection processes

Problem setting. Consider a group of N applicants applying for a job. The decision maker's goal is to hire the best possible βN applicants. The applicant n has an unknown quality level $p_n \in [0, 1]$. At each decision epoch t, the decision maker can ask $A = \{0, 1, 2\}$ questions to each candidate, and a signal $q_n(t) \sim \mathrm{binomial}(a, p_n)$ is returned, indicating how many among the a questions have been solved correctly by the applicant. Choosing an action a consumes $D(a)$ units of resource (time, space, organization cost, etc.), for which we choose $D(0) = 0, D(1) = 1$ and $D(2) = 1.5$. At each decision epoch a total amount of αN resource is available. There is a total number of T interviewing rounds, and in the final $(T + 1)$-th round, the decision maker admits βN applicants, based on the results of the interviewing rounds. We call the above problem the unfair case. We also consider a fair case where there are two groups of applicants (of size $N/2$) and for which the decision maker cannot spend more that γN budget on each of the single group, where $\gamma < \alpha < 2\gamma$.

These two problems can be cast into weakly coupled MDPs by considering a Bayesian model in which the quality level p of an applicant from each group is generated from some beta distribution, and the decision maker's estimation on each applicant is updated using Bayes' rule (see [4] for details). They generalize the application screening problem introduced in [2] by allowing more than two actions and by adding the fairness constraints.

Numerical results. We consider two scenarios. The first one is when the resource is "scarce". The second scenario where the resource is "abundant" where the resource is doubled. In each scenario we compare the effect of adding or removing the fairness constraints. Each time, we compare the performance of our LP-update policy, of the relaxed upper bound, and of a policy called the *occupation measure policy*, which is also proven to be asymptotically optimal in [8]. We report the results in Figure 1.

As guaranteed by theory, for all scenarios, both the LP-update policy and occupation measure policy converge to the LP-relaxed bounds, as N goes to infinity. Yet, the situation is different for smaller values of N: In all cases, the LP-update policy outperforms the occupation measure policy. The smaller the value of N, the more apparent is the advantage of the LP-update policy. More interestingly, we observe in the right panel of Figure 1 that for the abundant resource case, the fairness constraint does not play a role for the relaxed problem (because the constraints are never saturated). This figure shows that our LP-update policy also enjoys this property: the performance that we obtain is identical with or without fairness: our algorithm respects fairness at cost 0 for this case. This is not the case of the occupation measure policy.

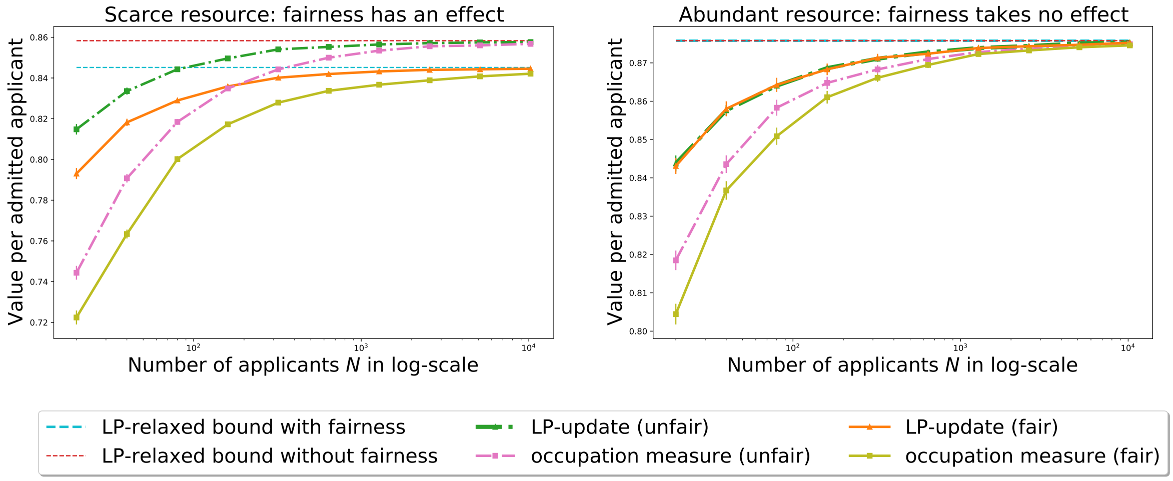

FIG. 1: Performance on generalized applicant screening problems when the resource is scarce (left panel) or abundant (right panel), with or without fairness constraints.

References

[1] Florin Avram, Dimitris Bertsimas, and Michael Ricard. Fluid models of sequencing problems in open queueing networks; an optimal control approach. *Institute for Mathematics and its Applications*, 71:199, 1995.

[2] David B. Brown and James E. Smith. Index policies and performance bounds for dynamic selection problems. *Manag. Sci.*, 66:3029–3050, 2020.

[3] Nicolas Gast, Bruno Gaujal, and Chen Yan. (close to) optimal policies for finite horizon restless bandits. *arXiv preprint arXiv:2106.10067*, 2021.

[4] Nicolas Gast, Bruno Gaujal, and Chen Yan. The lp-update policy for weakly coupled markov decision processes. *arXiv preprint arXiv:2211.01961*, 2022.

[5] Yasin Gocgun and Archis Ghate. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers and Operations Research*, 39(10):2323–2336, 2012.

[6] Jackson A Killian, Andrew Perrault, and Milind Tambe. Beyond" to act or not to act": Fast lagrangian approaches to general multi-action restless bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 710–718, 2021.

[7] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res*, pages 293–305, 1999.

[8] Guojun Xiong, Jian Li, and Rahul Singh. Reinforcement learning for finite-horizon restless multi-armed multi-action bandits. *arXiv preprint arXiv:2109.09855*, 2021.

[9] Xiangyu Zhang and Peter I Frazier. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv preprint arXiv:2203.15853*, 2022.