Incremental Improvements of Heuristic Policies for Average-Reward Markov Decision Processes

S. Reveliotis * **M. Ibrahim** **

* *School of Industrial & Systems Engineering*
Georgia Institute of Technology, USA
(email: spyros@isye.gatech.edu)
** *Computer Engineering Department*
Cairo University, Egypt
(e-mail: maykelnawar@hotmail.com)

Abstract: Within the realm of Discrete Event Systems (DES) theory, the problem of performance optimization for many applications can be modeled as an infinite-horizon, average-reward Markov Decision Process (MDP) with a finite state space. In principle, these MDPs can be solved by various well-developed methods like value iteration, policy iteration and linear programming. But in reality, the tractability of these methods in the context of the aforementioned applications is compromised by the explosive size of the underlying state spaces, a problem that is known as "the curse of dimensionality". Hence, the corresponding performance optimization problems are frequently addressed by heuristic control policies. The considered work uses results from (i) the sensitivity analysis of Markov reward processes and (ii) the ranking & selection theory in statistics in order to develop a methodology for assessing the optimality of isolated decisions in the context of any well-defined heuristic control policy for the aforementioned MDPs. It also determines an improved decision when the current one is found to be suboptimal. Hence, when embedded in an iterative scheme, this methodology can support the incremental enhancement of the original heuristic policy in a way that controls, both, the computational and also the representational complexity of the new policy. Finally, an additional important feature of the presented methodology is that it can be executed either in an "off-line" mode, using a simulation of the dynamics of the underlying DES, or in an "on-line" mode, based on the sample path that is defined by the real-time dynamics of the controlled system.

Keywords: Markov Decision Processes, Performance Optimization, Sensitivity Analysis of Markov reward processes, Ranking & Selection, Discrete Event Systems

1. INTRODUCTION

Finite state-space, infinite-horizon, average-reward Markov decision processes (AR-MDPs) (Puterman (1994)) is a natural formal framework for modeling many performance control and optimization problems that are formulated in the operational context of many contemporary applications. Under certain structural conditions that must be met by their underlying state space, these MDPs can be solved, at least in principle, for an optimal control policy by a host of well-developed methods, like value iteration, policy iteration and linear programming (Puterman (1994)). However, the computational tractability of all these methods is severely limited by the facts that (a) they employ a complete enumeration of the underlying state space, and (b) the cardinality of these state spaces usually increases very fast to some very large values. A natural explanation for this state-space explosion can be based on the fact that the state space of the considered MDPs is typically obtained as the product of the state spaces that correspond to the various components of the underlying plant system (Cassandras and Lafortune (2008)); therefore, this explosion has been characterized as the "curse of dimensionality" in the corresponding literature (Bellman (1957)). Furthermore, an additional remark that is important for the developments that are presented in this work, is that this "curse of dimensionality" effect does not imply only a prohibitive cost for computing an optimal policy for the considered MDPs, but also a very high *representational* cost for these policies, since, according to their basic definition, they need to specify an optimized decision rule for each state.

In view of the representational and computational complications that are indicated in the previous paragraph, many of the corresponding performance control and optimization problems are eventually addressed through an approximating or heuristic solution method that provides a suboptimal but more tractable policy. Hence, for instance, a popular methodology in the emerged field of approximate dynamic programming (ADP – Bertsekas (2012); Powell (2007)) has tried to address the aforementioned complications by employing a continuous approximation of the relative value function that defines the sought policy; this approximation is defined by means of a set of "basis

functions" that try to capture important operational characteristics and nonlinearities of the underlying system.

Some other methods that have also been considered by the current ADP theory are those known as "approximation in the policy space" and "state-space aggregation". The first of these methods essentially predefines a parameterized policy space, and tries to compute an optimized policy in this restricted policy space by carefully selecting the corresponding set of parameters; this parameter selection can be formulated as an optimization problem that usually is solved through techniques borrowed from simulation-based optimization and stochastic approximation (Asmussen and Glynn (2007)). On the other hand, methods based on state-space aggregation essentially try to control the representational complexity of the target policies by identifying clusters of states that are expected to have a common (or "similar") optimal decision, and explicitly enforcing this decision commonality upon these clusters; an optimized set of such common decisions is eventually computed through techniques similar to those used by the "approximation in the policy space" methods. [1]

An additional popular, but also powerful, approach for generating efficient suboptimal policies for the considered MDPs is that based on the notion of "fluid relaxation" (FR – Weiss (2000); Meyn (2008); Bertsimas et al. (2015); Ibrahim and Reveliotis (2019a)). FR-based methods consider a "fluidized" version of the dynamics of the underlying plant system, and at each decision point of the original system, they select an optimized decision for this point by (i) solving an optimal control problem that is formulated in the "fluidized" dynamics, and (ii) using the information that is contained in the optimal solution of this new problem as "guidance".

Finally, it is also true that, in certain cases, a pertinent policy for the considered MDP formulations can be determined in a quite *ad hoc* manner, on the basis of the existing analytical understandings and insights for the underlying performance optimization problem and its driving dynamics, and/or past empirical evidence.

Recognizing all the aforementioned realities regarding the solution of the performance control and optimization problems that are the focus of this work, and of the corresponding AR-MDP formulations, in this work we seek the development of a theoretical framework and effective computational tools that can assess the quality of the decisions that are effected by any given heuristic policy for some considered AR-MDP formulation, and also recommend potential improvements to these decisions if they are found to be suboptimal. Furthermore, taking into consideration (i) the representational and computational complications that result from the "curse of dimensionality" that haunts the targeted applications, and also (ii) the overall efficiency of the policies that are returned by the currently employed approaches, with respect to these complications, we want to take advantage of the efficiencies that are established by these initial policies, modifying them only in an incremental and localized manner. Finally, since the original heuristic policies usually will define a pretty good "baseline" for an effective and efficient operation of the underlying plant system, we also want our policy improving scheme to be implementable in real-time; in this way, the method to be presented in this work can also be perceived as a "learning mechanism" that can augment the performance of the underlying plant system by taking advantage of its own experiences.

From a methodological standpoint, the presented developments are enabled by results coming from (i) the area of the sensitivity analysis of Markov reward processes (Cao (2007)), and (ii) an area of statistical inference that is known as "Ranking & Selection" (R&S – Kim and Nelson (2006)). We shall introduce more systematically and review the corresponding results in the subsequent parts of the paper. Furthermore, the subsequent developments will also reveal that, at a higher level, this work shares a common methodological base with the policy iteration method for AR-MDPs; in fact, it can be perceived as a (very) "asynchronous" implementation of this method, in the corresponding terminology of Bertsekas (2012). When viewed from this viewpoint, another particular work that has a considerable methodological affinity with our developments, is that of Cooper et al. (2003). But instead of placing the focus on the convergence to an optimal policy, which is the main objective of any policy-iteration type of analysis, the considered work seeks to define the computational tools and methods that will attain the aforestated objective of the performance enhancement of some good heuristic policy that might be already available in the considered application context, while retaining representational and computational tractability.

From an application standpoint, the presented developments have been further motivated by – and tested in the context of – a particular application problem that has come to be known as the throughput maximization of capacitated re-entrant lines (CRLs). A first introduction of this problem can be found in (Reveliotis (2000); Choi and Reveliotis (2003)), while an FR-based solution approach to it is provided in the more recent publications of Ibrahim and Reveliotis (2019a,b); Ibrahim (2019). The work of Ibrahim (2019) also provides a series of experimental results regarding the testing of the methodology that is presented in this paper in that particular application context.

In view of the above positioning of the paper content and its intended contribution, the rest of it is organized as follows: The next section defines formally the class of the MDP formulations that are considered in this work, and introduces all the corresponding terminology and a set of additional assumptions that will facilitate the presented developments. These developments themselves are presented in Section 3, which states formally the particular problem that is addressed in this work, and details a solution approach for this problem. Section 4 provides some discussion that elaborates further on the presented developments in Section 3 and their applicability. Finally, Section 5 concludes the paper and suggests some directions for future work.

[1] Actually, it should be obvious to the reader that the methods based on state-space aggregation essentially constitute a particular class of "approximation in the policy space" methods, where the considered policy space and its parameterization are defined by the employed state aggregation scheme.

2. THE CONSIDERED MDP FORMULATIONS

The MDP formulations that are considered in this work can be represented by a tuple $\langle X, A, S, \mathcal{P}, R\rangle$ where the various components of this tuple are defined as follows:

(1) X is the set of *"decision states"* of the defined MDP. In this work, X is assumed to be finite.
(2) A is a finite set of *"decisions"* – or *"actions"* – that can be effected at the various decision points of the considered MPD. In particular, for each decision state x, there is a set $A_x \subseteq A$ that defines the set of possible actions that are available at x.
(3) S is the set of the *"post-decision states"* of the considered MDP. More specifically, the execution of a decision $a \in A_x$ at some state $x \in X$, results in a post-decision state $s(x,a) \in S$. State $s(x,a)$ is uniquely defined for every pair $(x,a) \in X \times A_x$. [2]
(4) $\mathcal{P}$ is a set of discrete distributions with common support set X. Each post-decision state $s \in S$ is associated with an element of the set $\mathcal{P}$ that will be denoted by P_s. The distribution P_s regulates the transition from the considered state s to the next decision state $x \in X$; hence, it will referred to as the *"distribution of the (one-step) transition probabilities"* from state s.
(5) R is the *"immediate reward function"* that associates with each post-decision state $s \in S$ an immediate reward value that will be denoted by $R(s)$.

It is clear from the above description that the considered MDP definition encompasses all of the typical instantiations of the finite-state-space MDPs that are considered by the classical MDP theory. The differentiation between the "decision state" x and the "post-decision state" s that is adopted for the representation of the underlying dynamics, is a frequently used convention (c.f. Bertsekas (2012)) and it will be useful in the presentation of the main results of Section 3.

In the subsequent developments, we shall also assume that the considered MDPs are *"communicating"*, i.e., there is a sequence of decisions that can lead from any state $q \in X \cup S$ to any other state $q' \in X \cup S$ with positive probability. In many practical applications, including the CRL throughput maximization problem that was discussed in the introductory section, this last assumption implies the possibility of averting problematic behavior like the system entrapment in deadlocks and livelocks, through the specification of a pertinent control policy. This possibility subsequently enables the specification of the performance objective to be attained by the sought control policies as the maximization of the average reward rate to be attained by the controlled MDP when operated over an infinite time horizon.

In more technical terms, let Π denote the set of the *stationary* control policies for the considered MDPs. Each element $\pi \in \Pi$ can be represented by a set of discrete probability distributions $P_x(\pi)$, $x \in X$, with corresponding support sets A_x. The specification of the policy π induces a discrete-time Markov chain (DTMC) on the corresponding space S, that will be denoted by $MC(\pi)$.

[2] Hence, the presumed finiteness of the sets X and A implies the finiteness of set S.

The communicating structure that is presumed for the considered MDPs implies that there exist policies $\pi \in \Pi$ for which the corresponding DTMC $MC(\pi)$ is ergodic (Puterman (1994); Bertsekas (2012)). In the following, we shall further restrict Π to denote the class of stationary policies π that result in an ergodic DTMC $MC(\pi)$, and we shall formally define the performance objective of the considered MDP formulations as

$$\max_{\pi\in\Pi} \eta(\pi) \equiv \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} R\big(s(x_t, a_t; \pi)\big) \tag{1}$$

The notation $s(x_t, a_t; \pi)$ in the above equation implies that the selection of the action a_t at each decision state x_t, $t = 0, 1, 2, \ldots$, is driven by the applied policy π. It can also be shown that, under the aforestated assumptions, there is an optimal policy $\pi^* \in \Pi$ for the objective of Eq. 1 that is deterministic; i.e., every distribution $P_x(\pi^*)$, $x \in X$, is selecting deterministically a single action of A_x, to be denoted by $a(x; \pi^*)$. In the following, we shall use the notation Π_d to denote the subset of Π that consists of deterministic policies, and we shall essentially focus on this particular class of stationary policies. But in order to guarantee that all the policies π that will be generated by the presented methodology will result in an ergodic DTMC $MC(\pi)$, we shall eventually consider a more randomized implementation of each policy $\pi \in \Pi_d$ according to the following scheme: At each decision state $x \in X$, this modified version of policy π will select action $a(x; \pi)$ with a high probability ζ; with the remaining probability $1 - \zeta$, the modified policy will select randomly an action $a \in A_x$ according to the uniform distribution. We shall denote by $\Pi_d^{(\zeta)}$ the set of policies that is induced from the policy set Π_d through this modification, and the policies to be considered in the following will be elements of the set $\Pi_d^{(\zeta)}$ for some large value ζ. Also, in order to avoid an overloading of the employed notation, in the following, we shall make the additional convention that any reference to a deterministic policy π will essentially imply its randomized counterpart in the considered set $\Pi_d^{(\zeta)}$.

Finally, for any policy $\pi \in \Pi_d^{(\zeta)}$, $\hat{P}(\pi)$ will denote the one-step transition probability matrix for the corresponding DTMC $MC(\pi)$, $\hat{\psi}(\pi)$ will denote the unique stationary distribution of $MC(\pi)$, and $\hat{\mathbf{g}}(\pi)$ will denote the relative value function – also, known as the *"potentials"* vector – for the Markov reward process that is induced by the DTMC $MC(\pi)$ and the immediate-reward function R. Then, letting (i) $\mathbf{r}$ denote a representation of the immediate-reward function R as a column vector with its components corresponding to the elements of the post-decision-state set S, (ii) I denote the identity matrix, and (iii) $\mathbf{1}$ denote the column vector with all of its components equal to 1.0, we also have (Puterman (1994)):

$$\eta(\pi) = \hat{\psi}(\pi)^T \cdot \mathbf{r} \tag{2}$$

$$\Big(I - \hat{P}(\pi)\Big) \cdot \hat{\mathbf{g}}(\pi) = \mathbf{r} - \eta(\pi)\mathbf{1} \tag{3}$$

3. MAIN RESULTS

The main problem considered in this work: In view of the definitions and the assumptions that were provided in the

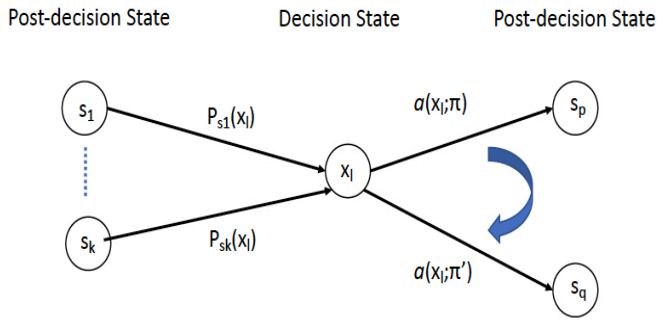

Fig. 1. A schematic representation of the main problem that is considered in this work.

previous section, the main problem to be considered in this work can be stated as follows:

Main problem Given a policy $\pi \in \Pi_d^{(\zeta)}$ and a decision state $x_l \in X$ from the underlying MDP, we want to develop a methodology that will assess the existence of an action $a \neq a(x_l;\pi)$ in A_x such that the local replacement of the action $a(x_l;\pi)$ by action a at state x_l will lead to a policy $\pi' \in \Pi_d^{(\zeta)}$ with $\eta(\pi') > \eta(\pi)$. In addition, we want the developed methodology to be able to address this issue under a single-sample-path-based analysis where the employed sample path is generated from the operation of the "plant" system under the original policy π.

In the rest of this section we show how to resolve the above problem using results from the sensitivity analysis of Markov reward processes and R&S type of algorithms. Also, the subsequent discussion will be facilitated by the schematic representation of the considered problem that is provided in Figure 1.

Sensitivity analysis of the considered MDPs: Consider two policies π and π' from the policy space $\Pi_d^{(\zeta)}$. Then, from Eq. 3 and the additional facts that

$$\hat{\psi}(\pi')^T \cdot \hat{P}(\pi') = \hat{\psi}(\pi')^T \tag{4}$$

and

$$\hat{\psi}(\pi')^T \cdot \mathbf{1} = 1.0 \tag{5}$$

we get:

$$\hat{\psi}(\pi')^T \cdot \Big((I - \hat{P}(\pi)\Big) \cdot \hat{\mathbf{g}}(\pi) = \hat{\psi}(\pi')^T \cdot \Big(\mathbf{r} - \eta(\pi)\mathbf{1}\Big) \Longrightarrow$$

$$\hat{\psi}(\pi')^T \cdot \Big(\hat{P}(\pi') - \hat{P}(\pi)\Big) \cdot \hat{\mathbf{g}}(\pi) = \hat{\psi}(\pi')^T \cdot \mathbf{r} - \eta(\pi)\Big(\hat{\psi}(\pi')^T \cdot \mathbf{1}\Big)$$

$$\Longrightarrow \quad \hat{\psi}(\pi')^T \cdot \Big(\hat{P}(\pi') - \hat{P}(\pi)\Big) \cdot \hat{\mathbf{g}}(\pi) = \eta(\pi') - \eta(\pi) \tag{6}$$

Equation 6 is a *"performance difference"* formula, characterizing the difference in the performance of the underlying MDP as we switch from policy π to policy π'. We can see that this difference is determined by (i) the elementwise difference of the one-step transition probability matrices $\hat{P}(\pi)$ and $\hat{P}(\pi')$ for the DTMCs $MC(\pi)$ and $MC(\pi')$ that are induced by these two policies, (ii) the relative value function $\hat{\mathbf{g}}(\pi)$ of policy π, and (iii) the stationary distribution $\hat{\psi}(\pi')$ of $MC(\pi')$.

When specialized to the policy pairs $(\pi, \pi') \in \Pi_d^{(\zeta)} \times \Pi_d^{(\zeta)}$ where policy π' is obtained from policy π through the single-decision modification that is depicted in Figure 1, the result of Eq. 6 takes the following form:

$$\eta(\pi') - \eta(\pi) = \zeta\Big(\sum_{s \in S} \hat{\psi}(s;\pi') \cdot P_s(x_l)\Big)\Big[\hat{\mathbf{g}}(s_q;\pi) - \hat{\mathbf{g}}(s_p;\pi)\Big] \tag{7}$$

Equation 7 implies that a policy modification of the type that is described in Figure 1 can result in an improvement of the performance of the underlying MDP if and only if

$$\hat{\mathbf{g}}(s_q;\pi) \;>\; \hat{\mathbf{g}}(s_p;\pi) \tag{8}$$

Hence, in order to effect a performance improvement for a currently running policy π through a policy modification of the type that is suggested in Figure 1, we need to identify a decision state x_l such that the decision under the current policy π corresponds to a post-decision state $s_p \equiv s(x_l, a(x_l;\pi))$ with $\hat{\mathbf{g}}(s_p;\pi) < \max_{a \in A_{x_l}} \hat{\mathbf{g}}(s(x_l,a);\pi)$. Eq. 7 further implies that the magnitude of the improvement that will result from the corresponding modification of the policy π, is determined by the difference $\hat{\mathbf{g}}(s_q;\pi) - \hat{\mathbf{g}}(\mathbf{s}_p;\pi)$, and it is further modulated by the other two factors that multiply this difference in Equation 7, i.e., the parameter ζ that defines the employed randomization for the considered policies, and the "steady-state" probability of visiting the considered decision state x_l under policy π'.[3]

Next we address the issue of developing pertinent estimators for the performance index $\eta(\pi)$ and the potentials $\hat{\mathbf{g}}(s;\pi)$, $s \in S$, for the Markov reward process that is induced by policy π, while the last part of this section presents a systematic methodology that will help us identify reliably improving modifications of the current policy π, based on these potential estimates.

Estimation of the performance index $\eta(\pi)$ and the state potentials $\hat{\mathbf{g}}(s;\pi)$, $s \in S$, for any given policy $\pi \in \Pi_d^{(\zeta)}$: Consider some arbitrarily chosen (post-decision) state $s^* \in S$. Then, it should be clear from the definition of the policies $\pi \in \Pi_d^{(\zeta)}$ that state s^* is a recurrent state for the DTMC $MC(\pi)$, and furthermore, every visit to the considered state s^* has a regenerative effect for the dynamics of $MC(\pi)$. But then, the performance index $\eta(\pi)$ can be computed by the formula

$$\eta(\pi) = \frac{E\Big[\sum_{t=0}^{\tau-1} \mathbf{r}\big(s(t)\big)\Big]}{E[\tau]} \tag{9}$$

where the random variable τ denotes the recurrence time for state s^* (Ross (1983)).

Equation 9 subsequently suggests the following estimator, $\widehat{\eta(\pi)}$, for the performance index $\eta(\pi)$: Simulate the DTMC $MC(\pi)$ initializing it to the selected state s^*, and let τ_N denote the time of the N-th recurrence of the process to state s^*. Then, set

[3] We also notice, for completeness, that while the state potentials $\hat{\mathbf{g}}(s;\pi)$, $s \in S$, provide good guidance for identifying improving policy modifications of the type described in Figure 1, when we try to assess the pertinence and the significance of such policy changes in an "off-line" mode, it is also possible to work more directly with estimates of the corresponding throughputs $\eta(\pi)$ and $\eta(\pi')$, obtained through simulation of the corresponding Markov reward processes.

$$\widehat{\eta(\pi)} \equiv \frac{\sum_{t=0}^{\tau_N-1} \mathbf{r}\big(s(t)\big)}{\tau_N} \tag{10}$$

The estimator $\widehat{\eta(\pi)}$ essentially employs the empirical means of the expectations $E[\tau]$ and $E\Big[\sum_{t=0}^{\tau-1} \mathbf{r}\big(s(t)\big)\Big]$ appearing in Equation 9, that are based on the N simulated recurrent cycles. Therefore, $\widehat{\eta(\pi)}$ is strongly consistent (i.e., $\widehat{\eta(\pi)} \to \eta(\pi)$ as $N \to \infty$ w.p. 1), but it is also biased for any finite N. In Asmussen and Glynn (2007) it is shown that the bias of $\widehat{\eta(\pi)}$ is $O(1/N)$, and if necessary, it can be further reduced to $O(1/N^2)$ by applying certain techniques like "jack-knifing". Furthermore, the $O(1/N)$ dependence of the bias of $\widehat{\eta(\pi)}$ on N implies that, when N takes fairly large values, this bias will be an order of magnitude smaller than the st. deviation of this estimator, which behaves as $O(1/\sqrt{N})$; therefore, the existing bias in the above estimator $\widehat{\eta(\pi)}$ is not expected to have a significant impact in the context of the developments that are presented in this work.

Next we shift our attention to the estimation of the state potentials $\hat{\mathbf{g}}(s;\pi),\ s \in S$. The presented method is based on the corresponding developments that appear in Cao (2007); Cooper et al. (2003). One way to motivate these developments is as follows: It is well known that a solution $\hat{\mathbf{g}}(\pi)$ for Equation 3 can be obtained by setting

$$\forall\, s \in S,\ \hat{\mathbf{g}}(s;\pi) = \lim_{L\to\infty} E\Big[\sum_{t=0}^{L-1} \Big(\mathbf{r}\big(s(t)\big) - \eta(\pi)\Big) \ \Big|\ s(0) = s\Big] \tag{11}$$

Furthermore, from Equation 11 it follows that for any state pair $(s_i, s_j) \in S \times S$ with $s_j \neq s_i$,

$$\gamma(s_i, s_j;\pi) \equiv \hat{\mathbf{g}}(s_j;\pi) - \hat{\mathbf{g}}(s_i;\pi) = E\Big[\sum_{t=0}^{\tau(i|j)-1} \Big(\mathbf{r}\big(s(t)\big) - \eta(\pi)\Big) \ \Big|\ s(0) = s_j\Big] \tag{12}$$

where

$$\tau(i|j) = \inf\{t > 0 : s(t) = s_i \mid s(0) = s_j\} \tag{13}$$

Finally, setting

$$\hat{\mathbf{g}}(s^*;\pi) = 0 \tag{14}$$

for some arbitrary state s^*, Equation 12 implies that the vector $\tilde{\mathbf{g}}(\pi)$ with

$$\tilde{\mathbf{g}}(s_i;\pi) = \begin{cases} 0, & \text{if } s_i = s^* \\ \gamma(s^*, s_i;\pi), & \text{o.w.} \end{cases} \tag{15}$$

is another valid solution for Equation 3.

Moreover, Equations 12 and 13 further imply that an estimator, $\widehat{\tilde{\mathbf{g}}(s_i;\pi)}$, of $\tilde{\mathbf{g}}(s_i;\pi)$, for any state $s_i \neq s^*$, can be obtained as follows: Consider a recurrent cycle of the DTMC $MC(\pi)$ with respect to the state s^*, of length τ. If this recurrent cycle does not visit state s_i, then it cannot provide an estimate of $\tilde{\mathbf{g}}(s_i;\pi)$. If, on the other hand, state s_i is visited during this recurrent cycle, then let τ_i denote the first period that state s_i is visited during this cycle. According to Equations 12 and 13, an estimate of $\tilde{\mathbf{g}}(s_i;\pi)$ is provided by

$$\widehat{\tilde{\mathbf{g}}(s_i;\pi)} = \sum_{t=\tau_i}^{\tau-1} \Big(\mathbf{r}\big(s(t)\big) - \widehat{\eta(\pi)}\Big) \tag{16}$$

In Equation 16, $\widehat{\eta(\pi)}$ is a previously obtained estimate of the throughput $\eta(\pi)$, for instance, through Eq. 10. Then, the estimator $\widehat{\tilde{\mathbf{g}}(s_i;\pi)}$ will be unbiased if the employed throughput estimate $\widehat{\eta(\pi)}$ is unbiased; otherwise, it will be biased. Finally, a more robust estimate of $\tilde{\mathbf{g}}(s_i;\pi)$ can be obtained by averaging the estimator of Equation 16 over N recurrent cycles with respect to state s^* that involve a visitation to the considered state s_i, for some appropriately selected value of N.

Detecting a performance-improving action at the considered decision state x_l through R&S theory: Next we discuss how to employ the derived estimators for the performance index $\eta(\pi)$ and the state potentials $\hat{\mathbf{g}}(s;\pi),\ s \in S$, in the context of some algorithmic procedures that will enable a robust comparison of the performance-improving potential of the various decisions $a \in A_{x_l}$, that are available at the considered decision state x_l, in spite of the noise that is present in these estimators. As already mentioned, these algorithmic procedures are provided by an area of statistical inference that is known as "ranking & selection" (Kim and Nelson (2006)); in the following, we provide a systematic description of the basic problem that is studied by the R&S theory, and we overview some further developments of this theory that are most relevant to this work.

The basic problem addressed by the R&S theory is the development of sampling processes that will enable the identification among a given set of "options", represented by the random variables $Y_1, \ldots, Y_k$, of an option Y_i that possesses the largest expected value. The satisfaction of this objective should be attained in a probabilistic and near-optimal sense. More specifically, the decision problems addressed by the R&S theory are further structured by the following assumptions:

Assumption 1 The performance measure of interest for each entertained option $i = 1, \ldots, k$, is the unknown mean μ_i of a *normal* random variable (r.v.) Y_i.

Assumption 2 Besides the means μ_i, the variances, σ_i, for the r.v.'s Y_i, are also unknown and possibly unequal.

Assumption 3 It is possible to generate a sequence $\langle \hat{Y}_{ij},\ j = 1, 2, \ldots \rangle$ of independent samples for each r.v. Y_i.

Assumption 4 The set of samples $\{\hat{Y}_{ij},\ i = 1, \ldots, k\}$ – i.e., the samples obtained for the r.v.'s $Y_i,\ i = 1, \ldots, k$, during the j-th round of sampling – can be independent or (positively) correlated.

Assumption 5 There is also a pre-specified parameter δ such that any pair of options $\{i, j\}$ with $|\mu_i - \mu_j| \leq \delta$ are treated as equivalent during the attempted comparison – in the corresponding terminology, the parameter δ defines an *"indifference zone"* for the pursued comparison.

Assumption 6 Finally, all the existing approaches also allow for an erring probability a with $a < 1/k$. [4]

The considered R&S problem itself can be stated as follows:

[4] The reader should notice that $1/k$ is the probability of selecting the correct option when this selection is performed completely randomly among the k available options; hence, Assumption 6 stipulates that any R&S method must perform better than the random selection.

R&S problem statement Under Assumptions 1-6, design a sampling process and the accompanying inference logic that will select an option i in the indifference zone of $\arg\max_{j\in\{1\ldots,k\}}\{\mu_j\}$ with probability $1-a$.

Of particular interest to this work are solutions to the aforestated version of the R&S problem that are *"fully sequential"*. Generally speaking, these solutions go through a first sampling round that enable them to collect some information about the inherent variability in each r.v. Y_i, and subsequently they perform an additional number of sampling rounds where the information that is obtained from the additional samples that are collected at each round is used to further assess the competitiveness of the options that are still entertained in that round, and potentially eliminate some of these options that are deemed to not be competitive anymore. The entire process terminates when the set of (remaining) competitive options becomes a singleton. These sequential procedures are especially useful to this work due to (i) their ability to make more expedient and efficient use of the information that is contained in the collected samples, eliminating early options that are not deemed competitive by this sampling process, and (ii) their implementability in a "real-time" operational setting.

Also, an additional potential feature of the R&S problem that was described in the previous paragraphs, is the presence of an option – to be denoted as the 0-th option – that is to be treated as the "preferred choice" as long as it belongs in the indifference zone of $\arg\max_{j\in\{0,1\ldots,k\}}\{\mu_j\}$. The corresponding version of the R&S problem is known as "comparison with a standard" in the relevant literature (Nelson and Goldsman (2001); Kim (2005)). This version is particularly relevant to this work, since we want to alter the current policy only if the expected gain from the contemplated modification(s) is significant.

Finally, a last remark concerns the extension of the R&S problem and the corresponding methodology to cases where the compared r.v.'s $Y_i,\ i=1,\ldots,k$, do not satisfy the normality requirement posed in Assumption 1. This extension can be attained by re-defining the j-th sample $\hat{Y}_{ij}$ of the i-th random variable Y_i as the average of N independent samples drawn from the corresponding distribution. Then, as long as the employed sample size N is adequately large, the central limit theorem (Ross (2014)) ensures that this modified sample concept follows approximately a normal distribution with mean $E[Y_i]$. Clearly, this sampling modification is very important for the application of the existing R&S algorithms to various practical settings, including those settings that constitute the focus of this work.

Indeed, for any given policy $\pi\in\Pi_d^{(\zeta)}$ and decision state $x\in X$, the problem that is addressed in this work can be framed as a R&S problem where the set of the available options is defined by the action set A_x, and corresponding r.v.'s Y_a, for $a\in A_x$, are the estimators $\tilde{\mathbf{g}}(\widehat{s(x,a)};\pi)$ that are defined through the selection of some state s^* for the specification of the regenerative cycles of the underlying process $MC(\pi)$. This last selection can be quite arbitrary, but from a computational standpoint, it is advantageous to select state s^* among the most visited states under policy π. Each regenerative cycle of the process $MC(\pi)$ with respect to state s^* defines a "sampling round", providing a new sample for each of those states $s(x,a),\ a\in A_x$, that are visited during this cycle. Since they are generated by the same sample path, samples obtained during the same regenerative cycle will be correlated. On the other hand, the regenerative nature of the considered cycles implies that samples obtained during different cycles will be independent. The distribution corresponding to each estimator $\tilde{\mathbf{g}}(\widehat{s(x,a)};\pi)$ will not be normal, but the eventually utilized samples can be "normalized" through the averaging method that was discussed in the previous paragraph. Finally, as already noticed, in the operational context of the considered MDPs, the action $a(x;\pi)$ defines a natural "standard" that should be overruled only if the resulting gains in the performance of the underlying MDP are expected to be substantial. [5]

A fully sequential procedure that addresses the "comparison with a standard" version of the considered R&S problem under the aforestated Assumptions 1-6, is provided in Kim (2005). Furthermore, from all the previous discussion, it is clear that this procedure also defines a very effective tool for resolving the R&S problems that we need to address in this work. Therefore, we replicate this procedure in Figure 2, focusing on its particular configurations that are of interest to this work.

4. DISCUSSION

In this section first we report briefly on the findings from a set of experiments that tested the methodology of Section 3 on some MDPs corresponding to the CRL throughput maximization problem that was mentioned in the introductory section, and subsequently we discuss the potential embedding of this methodology in a search process that can enable an incremental improvement of any starting deterministic policy π while controlling the computational and the representational complexity of the derived policies π'.

The considered experiment and its major findings: In order to test the efficacy of the methodology of Section 3, and the intensity of the involved computations, we applied this methodology to a number of MDPs that correspond to the CRL throughput maximization problem of Ibrahim and Reveliotis (2019a). The particular CRL configurations that were used for the generation of these MDPs were the 20 configurations that were employed in some experiments reported in that work; for each of these configurations we generated 30 specific CRL instantiations by further selecting, randomly, the timing parameters for the involved processing stages, and we subsequently discretized the resulting dynamics through uniformization. Hence, in total, we considered 600 MDPs. The original policy π for each of these MDPs was the policy defined by the FR-based scheduling approach for the considered CRLs that is presented in Ibrahim and Reveliotis (2019a,b). And the particular decision state x_l that was assessed by our methodology, was the most

[5] Along these lines, it is also interesting to notice that since the two factors that multiply the difference $[\hat{\mathbf{g}}(s_q;\pi)-\hat{\mathbf{g}}(s_p;\pi)]$ in the right-hand-side of Eq. 7 can be interpreted as probabilities, this potential difference also defines an upper bound for the expected gain with respect to the performance index $\eta(\pi)$.

Setup: Select confidence level $1-a$, indifference-zone parameter δ, and the first-stage sample size n_0. Also, set

$$\beta = \begin{cases} 1-(1-a)^{1/k} & \text{if the obtained samples for the various options at each sampling iteration are independent} \\ a/k & \text{if the obtained samples for the various options at each sampling iteration are correlated} \end{cases}$$

and determine ρ by solving the equation

$$\big(1+2\rho\big)^{-(n_0-1)/2} = 2\beta$$

Initialization: Let $\mathcal{O} = \{0, 1, 2, \ldots, k\}$ be the set of options in contention. Obtain n_0 observations $\hat{Y}_{ij}, j = 1, 2, \ldots, n_0$, from each option $i = 0, 1, 2, \ldots, k$.
For all $i \neq l,\ l = 0, 1, 2, \ldots, k$, compute S^2_{il}, the sample variance of the difference between option i and option l, and let

$$a_{il} = \frac{\rho(n_0-1)S^2_{il}}{\delta_{il}} \qquad \text{and} \qquad \lambda_{il} = \frac{\delta_{il}}{2}$$

where

$$\delta_{il} = \begin{cases} \delta/2, & \text{if } i = 0 \ \vee l = 0 \\ \delta, & \text{o.w.} \end{cases}$$

Set the observation counter $r := n_0$ and go to *Screening*.
Screening: For each $i < l$, $i \in \mathcal{O}$ and $l \in \mathcal{O}$, if

$$\sum_{j=1}^{r} \big(\mathcal{Y}_{ij} - \mathcal{Y}_{lj}\big) \ \leq \ -\max\Big\{0,\ a_{il} - \lambda_{il} r\Big\},$$

then eliminate i from $\mathcal{O}$; else if

$$\sum_{j=1}^{r} \big(\mathcal{Y}_{ij} - \mathcal{Y}_{lj}\big) \ \geq \ \max\Big\{0,\ a_{il} - \lambda_{il} r\Big\},$$

then eliminate l from $\mathcal{O}$.
In the above inequalities,

$$\mathcal{Y}_{ij} = \begin{cases} \hat{Y}_{ij} + \delta/2, & \text{if } i = 0 \\ \hat{Y}_{ij}, & \text{o.w.} \end{cases}$$

Stopping Rule: If $|\mathcal{O}| = 1$, then stop and select the option whose index is in $\mathcal{O}$. Otherwise, set $r = r + 1$, take one additional observation, $\hat{Y}_{ir}$, from each option $i \in \mathcal{O}$, and go to *Screening*.

Fig. 2. The fully sequential procedure of Kim (2005) for resolving the "comparison with a standard" version of the R&S problem under Assumptions 1–6. It is assumed that the "standard" value μ_0 is unknown, and the parameter c that appears in the deliberations of Kim (2005), has been set equal to 1, according to the corresponding recommendations that are provided in that work.

visited state under the control of policy π, according to a generated sample path that consisted of 100,000 regenerative cycles. The parameter δ specifying the size of the indifference zone for the R&S algorithm was set to $0.01\tilde{\mathbf{g}}(s(x_l, \widehat{a(x_l;\pi)});\pi)$ and $0.001\tilde{\mathbf{g}}(s(x_l, \widehat{a(x_l;\pi)});\pi)$, where the value of $\tilde{\mathbf{g}}(s(x_l, \widehat{a(x_l;\pi)});\pi)$ was that obtained from the 100,000 regenerative cycles that were mentioned above. Finally, the values for the erring probability a employed by the R&S algorithm were 0.05 and 0.01.

Under the aforementioned parameterizations, the R&S algorithm of Figure 2 was always able to select correctly an action $a \in A_{x_l}$ with a potential $\tilde{\mathbf{g}}(s(x_l, a(x_l;\pi));\pi)$ that belonged in the specified indifference zone. At the same time, our experiments indicate that the size of the indifference zone δ can impact drastically the amount of sampling that is necessary for the algorithm of Figure 2 in order to reach a decision; in particular, increasing the size of the indifference zone δ from $0.001\tilde{\mathbf{g}}(s(x_l, \widehat{a(x_l;\pi)});\pi)$ to $0.01\tilde{\mathbf{g}}(s(x_l, \widehat{a(x_l;\pi)});\pi)$ reduced this required amount of sampling by some orders of magnitude. The space limitations imposed on this work do not allow a more detailed presentation of the considered experiment and the obtained results; but a more complete account of this experiment, together with an extensive tabulation of the obtained results, can be found in Ibrahim (2019).

Extending the presented developments into a policy-improving mechanism: Next, we provide a few remarks on how to extend the developments that were presented in Section 3 of this paper, into a complete mechanism able to support the systematic improvement of the deterministic policies π that have been considered in this work.

The proposed policy-improving mechanism is essentially an iterative search process. Starting with the original heuristic policy, at each iteration the proposed search scheme will select one or more decision states and it will assess the efficacy of the decisions that are effected at those states by the current policy π. More specifically, the selection of a set $\tilde{X} = \{x_1, x_2, \ldots, x_n\}$ of decision states to be assessed at a given iteration, defines a set of potential modifications to the current policy π which is succinctly described by the set $A_{x_1} \times A_{x_2} \times \ldots \times A_{x_n}$; each element of this set is an n-tuple of decisions that defines a modified policy π' through the replacement of the decision of policy π at each state $x_i \in \tilde{X}$ with the corresponding decision that is contained in this tuple. Collectively, all policies π' that are defined in this way define a "local neighborhood" of policy π that must be searched for the best policy. The local search itself can be based on the simultaneous application of the R&S algorithm of Figure 2 on all states $x_i \in \tilde{X}$.

The selection of the decision states to be included in the aforementioned sets $\tilde{X}$ will affect the performance of this search scheme. Eq. 7 in Section 3 seems to suggest that the decision states most visited by the current policy π constitute good candidates for the considered sets $\tilde{X}$. But it is also true that if the starting heuristic policy is of high quality, then, these most visited states might be less likely to generate actual improvements for the current policy, since the corresponding decisions might be already optimized. On the other hand, improving the decision at a decision state that is visited with a low frequency by an optimized policy might not have a major impact on the overall performance of the underlying system, to the point that it might not be worth considering. In certain cases, the definition of the policy π itself might also suggest classes of decision states of the underlying MDP where the policy decisions are expected to be suboptimal. The bottom line is that we need some further analysis and methodology that will rationalize and systematize

the construction of the sets $\tilde{X}$ at each iteration of the contemplated search process.

On the representational side, the considered mechanism will need to explicitly store the identified decisions that will differ from the decisions that are effected by the original heuristic policy. If the original policy is of good quality, then, the decision points where this policy will need to be corrected might not be that many; this is especially true when we also take into consideration the above discussion about the potential insignificance of less visited states in the overall performance of the derived policy. But, more generally, the form and the amount of storage space that should be provided for tracing the effected modifications to the original policy π should be considered as an additional "parameter" of the presented scheme that should be determined before its implementation.

A last theme that can be further considered in the context of the proposed search mechanism, is the selection of the parameters a and δ for the employed R&S algorithm, and of the parameter ζ that determines the extent of randomization in the implementation of the considered policies. Clearly, the specification of some tight values for the parameters a and δ will establish a high discerning power for the resulting algorithm, and a high quality for the corresponding decisions. But the reported experiments in the previous part of this section indicate that a very tight value for the parameter δ will result in a large amount of sampling; hence, it might be pertinent to start the proposed search by using some more relaxed values for the parameter δ and tighten these values as the overall search for improving policies becomes more difficult. As for the pricing of the parameter ζ, this issue must be resolved by considering the difficulty of reaching the various post-decision states of interest during the MDP operation under any fixed policy $\pi \in \Pi_d^{(\zeta)}$. More specifically, if some of these states are not easily accessible under a particular realization of the policy π, then, it will take a large number of regenerative cycles in order to collect the necessary samples of the corresponding state potentials. This problem can be mitigated by decreasing the employed value of ζ; i.e., by increasing the randomization level in the implementation of π. But it is also true that smaller values for ζ will attenuate the original performance difference between any pair of deterministic policies π and π', rendering harder the comparison of these two policies.

5. CONCLUSIONS

This work has developed a methodology for detecting sub-optimal decisions in any heuristic policy π that might be available for a well-defined AR-MDP formulation, and for recommending potential modifications for these decisions that can lead to an enhanced performance for the underlying MDP. Out future work will seek to extend these developments in a full-fledged policy improving mechanism, along the lines that are discussed in the second part of Section 4.

Acknowledgement

This work was partially supported by NSF grant ECCS-1707695.

REFERENCES

Asmussen, S. and Glynn, P.W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, NY, NY.

Bellman, R. (1957). *Applied Dynamic Programming*. Princeton University, Princeton, N. J.

Bertsekas, D.P. (2012). *Dynamic Programming and Optimal Control, Vol. 2 (4th ed.)*. Athena Scientific, Belmont, MA.

Bertsimas, D., Nasrabadi, E., and Paschalidis, I.C. (2015). Robust fluid processing networks. *IEEE Trans. on Automatic Control*, 60, 715–728.

Cao, X. (2007). *Stochastic Learning and Optimization*. Springer, NY,NY.

Cassandras, C.G. and Lafortune, S. (2008). *Introduction to Discrete Event Systems (2nd ed.)*. Springer, NY, NY.

Choi, J.Y. and Reveliotis, S.A. (2003). A generalized stochastic Petri net model for performance analysis and control of capacitated re-entrant lines. *IEEE Trans. on Robotics and Automation*, 19, 474–480.

Cooper, W.L., Henderson, S.G., and Lewis, M.E. (2003). Convergence of simulation-based policy iteration. *Probability in Engineering and Information Science*, 17, 213–234.

Ibrahim, M. (2019). *Scheduling Techniques for Complex Resource Allocation Systems*. Ph.D. thesis, ISyE, Georgia Tech, Atlanta, GA.

Ibrahim, M. and Reveliotis, S. (2019a). Throughput maximization of capacitated re-entrant lines through fluid relaxation. *IEEE Trans. on Automation Science and Engineering*, 16, 792–810.

Ibrahim, M. and Reveliotis, S. (2019b). Throughput maximization of complex resource allocation systems through timed-continuous-Petri-net modeling. *Discrete Event Dynamic Systems: Theory and Applications*, 29, 393–409.

Kim, S.H. (2005). Comparison with a standard via fully sequential procedures. *ACM Trans. on Modeling and Computer Simulation*, 15, 155–174.

Kim, S.H. and Nelson, B.L. (2006). Selecting the best system. In S.G. Henderson and B.L. Nelson (eds.), *Handbook in Operations Research and Management Science: Simulation*. Elsevier.

Meyn, S. (2008). *Control Techniques for Complex Networks*. Cambridge University Press, Cambridge, UK.

Nelson, B. and Goldsman, D. (2001). Comparison with a standard in simulation experiments. *Management Science*, 47, 449–463.

Powell, W.B. (2007). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, NY, NY.

Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Reveliotis, S.A. (2000). The destabilizing effect of blocking due to finite buffering capacity in multi-class queueing networks. *IEEE Trans. on Autom. Control*, 45, 585–588.

Ross, S.M. (1983). *Stochastic Processes*. Wiley, N.Y.

Ross, S.M. (2014). *A First Course in Probability (9th edition)*. Pearson, N.Y.

Weiss, G. (2000). Scheduling and control of manufacturing systems – a fluid approach. In *Proceedings of the 37th Allerton Conference*, –. University of Illinois.