Optimal Node Visitation in Stochastic Digraphs

Theologos Bountourelis and Spyros Reveliotis

Abstract— **The *Optimal Node Visitation (ONV)* problem addressed in this paper concerns the visitation of a subset of nodes in a stochastic graph a specified number of times, while minimizing the expected visits to another node in this graph. The presented results first provide a formulation of the ONV problem as a *stochastic shortest path (SSP)* problem, and subsequently they develop a suboptimal policy that is computationally tractable and asymptotically optimal. In particular, it is established that the ratio of the expected performance of this policy to the expected performance of an optimal policy converges to one, as the underlying visitation requirements are scaled uniformly to infinity. Furthermore, it is shown that under some stronger assumptions, the divergence of the performance of this policy from the performance of the optimal policy remains uniformly bounded by a constant, as the visitation requirements are scaled to infinity. Finally, it is shown that, for certain problem structures, the considered policy admits a closed-form characterization of its performance, which subsequently enables its optimized paremeterization and its efficient integration into adaptive control schemes of even higher efficiency.**

Index Terms— **Markov Decision Processes, Stochastic Shortest Path Problems, Fluid Modeling, Suboptimal Control, Asymptotic Analysis**

I. Introduction

The problem addressed in this work can be briefly described as follows: A control agent moves through a set of nodes, X, by selecting at each visited node $x \in X$ an action, a, from a set $\mathcal{A}(x)$, that will transfer it to a node $x' \in \mathcal{S}(a) \subseteq X$, with probability $p(x';a)$. The agent is initially placed at a certain node $x_0 \in X$, and its objective is to visit each node $x^j \in X^T \subset X$, $j = 1,, 2 \ldots, |X^T|$, a certain number of times, $\mathcal{N}_j$, while minimizing, in expectation, the number of its visits to a particular node $x^0 \in X \backslash X^T$. A preliminary study of this problem was presented in [1], where it was further assumed that (i) $x_0 = x^0$, (ii) the transitions out from the target nodes $x^j \in X^T$ lead deterministically to node x^0 and (iii) the subgraph obtained from the elimination of the transitions mentioned in (ii) has an *acyclic* structure with node x^0 defining its unique *"root"* node and the target nodes $x^j \in X^T$ being a subset of the *"terminal"* nodes. In [1], it was shown that this more restricted version of the problem accepts a *stochastic shortest path (SSP)* formulation [2], but that the underlying state space increases exponentially with respect to $|X^T|$. Hence, the work of [1] also introduced a sub-optimal randomized policy that was defined on the basis of a continuous – or *"fluid"* [3], [4], [5]– relaxation of the original problem, and it was shown to be *asymptotically optimal*, in the sense that the ratio of its performance to the performance of an optimal policy converges to one, as the node visitation requirements are scaled uniformly to infinity. The work presented in this paper seeks to extend and strengthen the results of [1] by (i) expanding them to address more general transition structures, (ii) enriching and improving the class of asymptotically optimal policies available for the ONV problem, and (iii) deriving stronger convergence properties for the performance of these policies.

More specifically, in this work we study the ONV problem without the restrictive assumptions on the connectivity of the underlying state space that were introduced in [1]. We show that under some very general assumptions that will guarantee the existence of *proper* policies [2], the problem retains its basic SSP structure, and there exists a *fluid-based* relaxation such that any optimal solution of the relaxed formulation induces an asymptotically optimal randomized policy. In the following, we shall refer to any such randomized policy that is induced by the aforementioned relaxed formulation, as policy π^{rel}. By using concepts and results borrowed from *renewal theory*, and especially the *central limit theorem (CLT)* for renewal processes [6],[1] we are also able to provide a bound for the rate of increase of the performance difference between policy π^{rel} and any optimal policy π^*, as the target node visitation requirements are scaled uniformly to infinity. Even more interestingly, this analysis has led to the identification of further conditions, of considerable generality, under which the asymptotic optimality of π^{rel} becomes stronger; in particular, under these additional conditions, the difference between the performance of π^{rel} and the performance of any optimal policy π^* remains uniformly bounded by a constant, as the node visitation requirements grow uniformly to infinity. Another part of the presented work discusses the possibility for more efficient *adaptive* implementations of π^{rel}, that take advantage of the special structure that is present in the state space of the aforementioned SSP formulation. More specifically, these adaptive implementations of π^{rel} retain its computational efficiency while they exploit the additional information provided in the "history" of the attained visitation requirements in order to achieve enhanced performance, by properly adjusting the action selection probabilities every time that a new visitation requirement is met. Finally, we also establish the rather surprising result that, under some additional assumptions on the underlying transition structure that subsume the topology studied in [1], the performance of policy π^{rel} can be characterized in *"closed-form"*. This result stems from our ability to characterize the underlying process dynamics through a Markovian scheme [6], [7], and, when applicable, it enables (i) an *optimized* implementation of π^{rel} on the considered ONV variations, and (ii) the efficient implemen-

The authors are with the School of Industrial & Systems Engineering, Georgia Institute of Technology, {tbountou,spyros}@isye.gatech.edu

This work was partially supported by NSF grants DMI-MES-0318657 and CMMI-0619978.

[1] instead of the *strong law of large numbers (SLLN)* that was originally used in [1]

tation of adaptive *"rollout"* policies [8] that employ π^{rel} as the underlying "base" policy.

From a methodological standpoint, the presented analysis for the ONV problem is similar, in spirit, to the prevailing trends regarding the analysis of stochastic scheduling problems [5], [9]. As indicated in [5], most stochastic scheduling problems are notoriously hard to solve optimally, and one has to compromise for solutions that are suboptimal but computationally tractable. In particular, the last few years have seen the emergence of a number of works that seek to provide suboptimal solutions to various stochastic scheduling problems by exploiting some "relaxed" – or "fluid"-based – version of the original problem [10], [11], [12]. Furthermore, in many cases, this line of analysis also provides guaranteed bounds for the potential suboptimality of the derived policies; cf., for instance, the works of [3], [4] and the references provided therein.

Finally, the practical motivation for the formulation and study of the ONV problem has been provided by our work presented in [13]. In that work, a learning agent must compute on-line an optimal policy for a task that evolves episodically over a state space that is stochastic and acyclic, and it has a single source state that defines the task initial state. It is shown that the agent can obtain an ϵ-optimal policy with probability at least $1-\delta$, by sampling the various actions available at each state a certain number of times[2] and selecting the action that results to the highest sample mean. Furthermore, this sampling must be performed on a layer by layer basis, starting from the terminal states and proceeding towards the initial state of the underlying state space. Higher-level states that have covered all the required sampling, and have their actions selected, are declared "fully explored", and they abandon the layer of "actively explored" states. On the other hand, lower-level states join the layer of "actively explored" states when all their immediate successors become fully explored. It is clear from the above that, in the considered setting, expedient learning translates to the determination of routing policies that will enable the realization of the required sampling in a minimum number of episodes. Furthermore, under the assumption that the transition probabilities are known *a priori*, the problem of determining such an optimized routing policy for a given set of actively explored states corresponds to the ONV problem variation defined in [1]. Hence, the ONV problem constitutes a *prototypical abstraction* whose study can offer the analytical insights and effective policies that subsequently can be implemented in the context of the learning algorithm described above, according to a *"certainty equivalence"* scheme [2] that substitutes the actual transition probabilities with pertinent estimates obtained during the execution of the algorithm. Another potential application context for the ONV problem variation defined in [1] is provided by various experimental setups where the subject must be studied in a number of states that are obtained from an initial state through some sequential treatment with probabilistic outcomes at the various stages. Assuming that the performed treatment has a destructive effect on the subject, one would like to obtain the required measurements while minimizing the number of subjects utilized in the experiment. Finally, for the extended version of the ONV problem studied in this work, one can easily envision additional "patrolling" and/or "roaming" applications over stochastically traversed "terrains" where an agent tries to visit some selected target areas a pre-specified number of times, while minimizing its exposition to a certain dangerous region.

Given the above positioning of the paper results, the rest of it is organized as follows: Section II provides the formal definition of the ONV problem considered in this work, and its further abstraction to an SSP problem. Section III introduces a further transformation ("reduction") of the problem that is necessary for the subsequent definition of the proposed asymptotically optimal policies. Section IV introduces π^{rel}, the primary policy considered in this work, and it formally establishes its asymptotic optimality. Section V addresses the potential of more elaborate, adaptive implementations of π^{rel}, with enhanced performance compared to the original policy. Section VI presents a series of additional results that pertain to the restricted ONV problem version studied in [1], and concern the capability of closed-form performance evaluation of π^{rel} and its implications. This section also reports a computational study that exemplifies and highlights the major analytical developments of the manuscript. Finally, Section VII concludes the paper and discusses directions for potential extensions of the presented results.

II. Problem description and its SSP formulation

A formal description of the ONV problem The ONV problem described in the introductory section is completely defined by a 7-tuple $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$, where:

- X is a finite set of *nodes* such that $\{x_0, x^0\} \subseteq X$ and $X^T = \{x^1, x^2, \ldots, x^{|X^T|}\} \subseteq X \backslash \{x^0\}$.
- $\mathcal{A}$ is a set function defined on X, that maps each $x \in X$ to the finite, non-empty set $\mathcal{A}(x)$, comprising all the *decisions / actions* that can be executed by the control agent at node x. It is further assumed that for $x \neq x'$, $\mathcal{A}(x) \cap \mathcal{A}(x') = \emptyset$.
- $\mathcal{P}$ is the *transition function*, defined on $\bigcup_{x \in X} \mathcal{A}(x)$, that associates with every action a in this set a discrete probability distribution $p(\cdot; a)$. The support sets of the distributions $p(\cdot; a)$ are subsets of the node set X, to be denoted by $\mathcal{S}(a)$, and it is further assumed that: (i) for every node $x \in X$, there is an action selection scheme – or a *policy* π – that renders x^0 accessible from x,[3] and (ii) for every node $x^j \in X^T$, $j = 1, 2, \ldots, |X^T|$, there is a policy π that renders x^j accessible from x^0.
- $\mathcal{N}$ is the *visitation requirement vector*, that associates with each node $x^j \in X^T$ a visitation requirement $\mathcal{N}_j \in \mathbb{Z}^+$.
- Finally, in order to facilitate the following discussions on the computational complexity of the ONV problem and the

[2] that depends on the graph structure and the performance parameters ϵ and δ

[3] i.e., starting from node x and selecting actions at each node according to policy π, there is a positive probability that the agent will reach node x^0

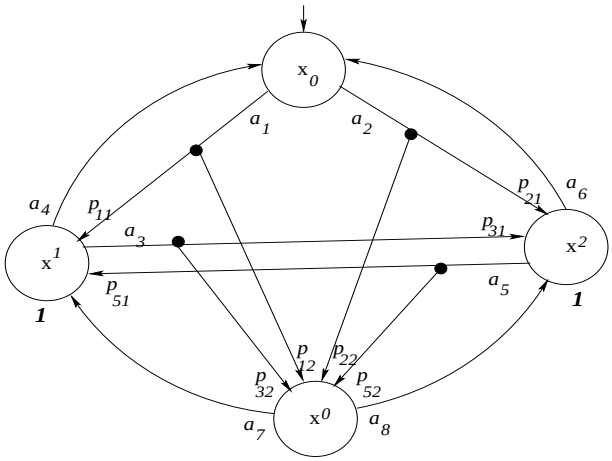

Fig. 1. An example problem instance

proposed solutions, we define the *instance size* $|\mathcal{E}| \equiv |X| + |\bigcup_{x\in X} \mathcal{A}(x)| + |\mathcal{N}|$, where application of the operator $|\ |$ on a set returns the cardinality of this set, while application on a vector returns its l_1 norm.

In the subsequent discussion we shall also employ the variable vector $\mathcal{N}^c$ to denote the *vector of the remaining visitation requirements*. The control agent starts from node x_0 at period $t = 0$, sets $\mathcal{N}^c := \mathcal{N}$, and at every consecutive period $t = 1, 2, 3, \ldots$, it (i) observes its current position, x_t, and the vector of remaining visitation requirements, $\mathcal{N}^c$, (ii) selects and executes an action $a \in \mathcal{A}(x_t)$, and (iii) upon reaching one of the target nodes, $x^j \in X^T$, updates $\mathcal{N}_j{}^c$ to $(\mathcal{N}_j{}^c - 1)^+$. The entire operation terminates when all the node visitation requirements have been satisfied, i.e., $\mathcal{N}^c$ has been reduced to zero. Our intention is to *determine a policy π that maps each tuple $(x, \mathcal{N}^c)$ to an action $\pi(x, \mathcal{N}^c) \in \mathcal{A}(x)$ in a way that will enable the agent to satisfy all the visitation requirements expressed by $\mathcal{N}$, while minimizing the expected number of visitations to node x^0.*

Example Figure 1 demonstrates the above definitions, by presenting a specific instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ of the considered ONV problem. In the depicted problem instance, the control agent is initially located to node x_0, and the set of target nodes is $X^T = \{x^1, x^2\}$, with respective visitation requirements $\mathcal{N}_1 = \mathcal{N}_2 = 1$. On the other hand, node x^0 is the node to be avoided during the execution of the requested visits to the aforementioned target nodes. It can also be noticed that $\mathcal{A}(x_0) = \{a_1, a_2\}$, $\mathcal{A}(x^1) = \{a_3, a_4\}$, $\mathcal{A}(x^2) = \{a_5, a_6\}$, $\mathcal{A}(x^0) = \{a_7, a_8\}$, and that the transitions resulting from these actions satisfy the connectivity requirements posed on the transition function $\mathcal{P}$; i.e., there are paths of positive probability from each node $x \in X$ to node x^0 and also paths that lead with positive probability from node x^0 to each of the two target nodes x^1 and x^2.

The induced stochastic shortest path problem The ONV problem described in the previous paragraph can be further abstracted to a *Discrete Time Markov Decision Process (DT-MDP)*, $\mathcal{M} = (S, A, t, c)$, such that:

- S is the finite set of *states*, identified with the tuples $(x, \mathcal{N}^c)$, where $x \in X$ and $\mathcal{N}^c \in \prod_{j=1}^{|X^T|}\{0, \ldots, \mathcal{N}_j\}$;
- A is a set function that associates with every state $s = (x, \mathcal{N}^c)$ the *action* set $A(s) \equiv \mathcal{A}(x)$, where $\mathcal{A}(x)$ is specified in the definition of $\mathcal{E}$;
- $t : S \times \bigcup_{s\in S} A(s) \times S \longrightarrow [0, 1]$, is the MDP *state transition* function that is induced from the above definitions of S and A, and the transition function $\mathcal{P}$ that appears in the definition of $\mathcal{E}$; and
- c is the *cost function* defined on S with

$$c(s) = \begin{cases} 1, & \text{if } x = x^0 \text{ and } \mathcal{N}^c \neq \mathbf{0}; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

It should be clear from the above specification of $\mathcal{M}$ that the set of states $s = (x, \mathcal{N}^c)$ with $\mathcal{N}^c = \mathbf{0}$ constitute a *closed* class which is also *cost-free*, i.e., once the process enters this class of states it will remain in it, and there will be no further cost accumulation. Hence, for the purposes of the subsequent developments, it is pertinent to aggregate this entire class of states into a single aggregate state, s^T, that is *absorbing* and *cost-free* under any policy π; we shall refer to the state s^T as the problem *terminal state*. Figure 2 exemplifies the resulting structure by depicting the state transition diagram for the MDP $\mathcal{M}$ that is induced by the ONV problem instance $\mathcal{E}$ depicted in Figure 1.

In the following, we are especially interested in a policy π^* that, starting from the *initial state* $s^0 \equiv (x_0, \mathcal{N})$, will drive the underlying process to the terminal state s^T with the minimum expected total cost. Let $V_\pi(s^0) = E_\pi[\sum_{t=0}^{\infty} c(s_t)|s_0 = s^0]$, where π is some given policy from the policy set Π, and the expectation $E_\pi[\cdot]$ is taken over all possible realizations under π. Then π^* is formally defined by

$$\pi^* = \arg\min_{\pi\in\Pi} V_\pi(s^0) \tag{2}$$

It is easy to see that the resulting MDP formulation is a well-defined *Stochastic Shortest Path (SSP)* problem [2]. Therefore, according to [2]:

Theorem 1: There exists a unique vector $V^*(s)$, $s \in S$, with $V^*(s^T) = 0$ and with its remaining components satisfying the Bellman equation

$$V^*(s) = \min_{a\in A(s)} \{c(s) + \sum_{s'\in S} t(s, a, s') \cdot V^*(s')\} \tag{3}$$

Furthermore, the vector $V^*(s)$ defines an optimal policy π^*, by setting for all $s \in S\backslash\{s^T\}$,

$$\pi^*(s) := \arg\min_{a\in A(s)} \{c(s) + \sum_{s'\in S} t(s, a, s') \cdot V^*(s')\} \tag{4}$$

The vector V^* introduced in Theorem 1 is known as the *optimal value function* or the *optimal cost-to-go vector* for the considered MDP formulation, since each component $V^*(s)$ expresses the expected total cost of initiating the underlying process at state $s \in S$ and subsequently following an optimal policy. Furthermore, Equation 4 implies that the availability of V^* enables the straightforward determination of an optimal policy. Yet, from a practical computational standpoint, the value of Theorem 1 in the determination of an optimal policy for any given problem instance,

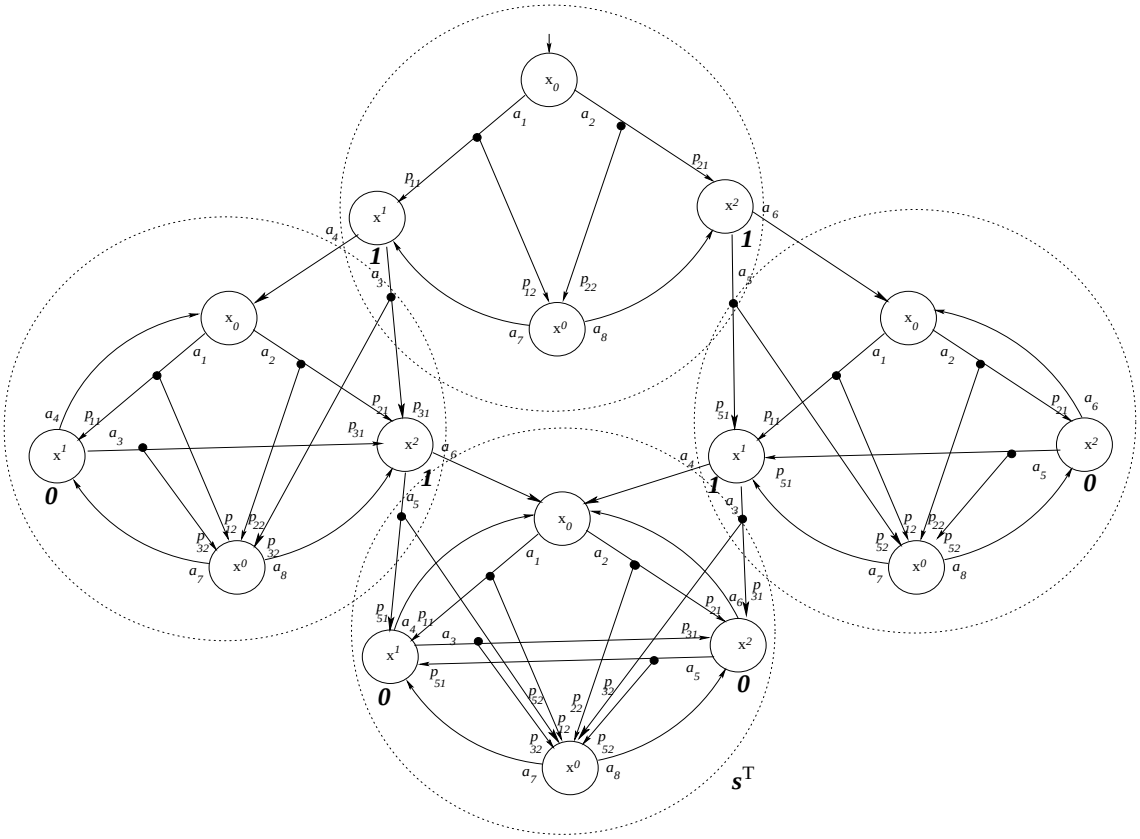

Fig. 2. The Stochastic Shortest Path problem corresponding to the ONV problem instance depicted in Figure 1

$\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$, is limited by the fact that the size of the state space, S, of the induced SSP problem grows exponentially with respect to the number of the problem target nodes, $|X^T|$, since $|S| = |X| \cdot \prod_{j=1}^{|X^T|}(\mathcal{N}_j+1) - |X| + 1$. On the other hand, the monotonic decrease of $\mathcal{N}^c$, and the acyclic structure in the underlying state space that is implied by this effect, enable the incremental solution of the formulation of Theorem 1 through a series of subproblems that are defined on the subspaces obtained by fixing the value for the remaining visitation requirement vector $\mathcal{N}^c$.[4] Clearly, each of these subproblems will be of polynomial complexity with respect to $|\mathcal{E}|$. But the set of all possible values for $\mathcal{N}^c$ is an exponential function of $|X^T|$, and therefore, the complexity of this decomposing approach remains super-polynomial.

Motivated by the observations of the previous paragraph, in the remaining part of this work we develop a number of suboptimal policies for the considered ONV problem that seek to trade off some operational efficiency for computational tractability. However, all the presented policies possess *asymptotic optimality*, in the sense that the *ratio* of their expected performance to the expected performance of the optimal policy converges to unity as the node visitation requirements grow uniformly to infinity. Furthermore, under some additional assumptions, we establish the even stronger result that the *difference* of the expected performance of these policies from the expected performance attained by any optimal policy will be uniformly bounded by a constant, as the visitation requirements are scaled uniformly to infinity. The detailed definition and implementation of the aforementioned suboptimal policies for the ONV problem necessitates its transformation (preprocessing) to an equivalent version where the underlying stochastic graph presents some additional structural properties. We shall refer to this transformed version of any given ONV problem as the *"reduced"* problem. The next section motivates and describes this reduction.

[4]For the example of Figure 2, these subspaces are indicated by the dotted circles.

III. The reduced ONV problem

The proposed reduction of the considered ONV problem is motivated by the following observation: Suppose that there exists a node subset $X' \subseteq X \backslash \{x^0\}$ and for every $x \in X'$ there are action subsets $\mathcal{A}'(x) \subseteq \mathcal{A}(x)$ such that the subgraph $\mathcal{G}'$ induced by X' and $\bigcup_{x \in X'} \mathcal{A}'(x)$ is closed and communicating. Then, a single access by the control agent of this subgraph through any node $x \in X'$ can guarantee the satisfaction of all the visitation requirements for all target nodes $x \in X^T \cap X'$. Hence, in the computation of any solution of the considered ONV problem, it will be pertinent to aggregate the subgraph $\mathcal{G}'$ into a single node with a visitation requirement of 1, if $x \in X^T \cap X' \neq \emptyset$, and 0 otherwise. Furthermore, it should not be difficult to see that for every node $x \in X \backslash \{x^0\}$, there is a unique maximal closed and communicating subgraph $\mathcal{G}'(x)$ pre-

Input: An ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$
Output: The reduced problem instance $\mathcal{E}^r = (X^r, x_0^r, x^0, X^{T^r}, \mathcal{A}^r, \mathcal{P}^r, \mathcal{N}^r)$

Stage 1
- $\mathcal{G}^0 := \mathcal{G}$ (where $\mathcal{G}$ denotes the stochastic graph corresponding to $\mathcal{E}$).
- Remove from $\mathcal{G}^0$ node x^0 and all the actions emanating from and leading to it.
- Repeat:

– Identify in $\mathcal{G}^0$ a node x with no actions emanating from it, and remove it from $\mathcal{G}^0$ as well as all the actions leading to it,

until the entire graph $\mathcal{G}^0$ has been eliminated or no node x can be identified.
- If the entire graph $\mathcal{G}^0$ has been eliminated, go to Stage 3; otherwise, proceed with Stage 2.

Stage 2
- $\mathcal{G}^1 := \mathcal{G}^0$ (where $\mathcal{G}^0$ is the stochastic graph obtained from the execution of Stage 1).
- Repeat:

– Pick a node x in $\mathcal{G}^1$ and compute the subgraph $\mathcal{G}'(x)$ of $\mathcal{G}^1$ corresponding to the equivalence class of x;

– remove $G'(x)$ from $\mathcal{G}^1$,

until all of $\mathcal{G}^1$ has been eliminated.

Stage 3
- Use the results of Stages 1 and 2 in order to compile and return the graph $\mathcal{G}^r$ corresponding to the reduced problem version $\mathcal{E}^r$. In particular, replace, in the original graph $\mathcal{G}$, every subgraph $\mathcal{G}'(x)$ with more than one nodes, computed in Stage 2, by a single aggregate node x', and associate to this node x' a unit visitation requirement, if $\mathcal{G}'(x)$ contains any target nodes, and a zero visitation requirement, otherwise.

Fig. 3. An algorithm for computing the reduced version, $\mathcal{E}^r$, of any given ONV problem instance $\mathcal{E}$

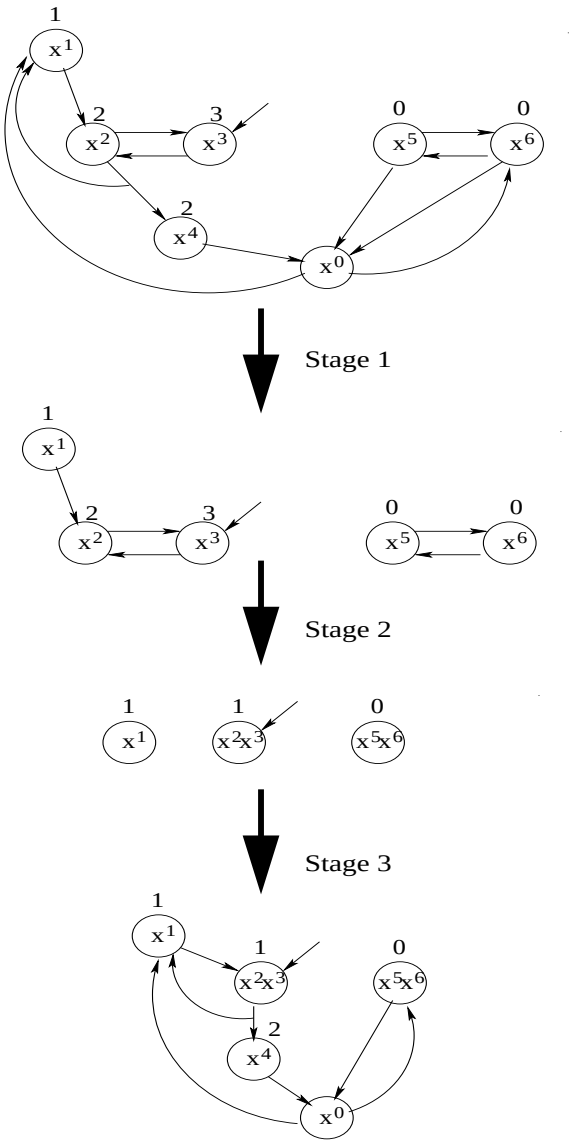

Fig. 4. Demonstrating the execution of the algorithm of Figure 3

sented the aforementioned properties,[5] and collectively, these maximal subgraphs define an equivalence relationship on $X\backslash\{x^0\}$. The reduced version of the ONV problem proposed in this section is obtained by aggregating the subgraph $\mathcal{G}'$ corresponding to each equivalence class of $X\backslash\{x^0\}$ by a single node. In this reduced problem representation, nodes corresponding to singleton equivalence classes preserve their visitation requirements while the remaining nodes possess a binary visitation requirement, evaluated as explained above. Finally, it is implicitly understood that whenever a node corresponding to a non-singleton equivalence class, with a unit visitation requirement, is visited for the first time, the agent will remain in the corresponding subgraph $\mathcal{G}'$ until all the visitation requirements for all the target nodes in this class have been covered.

An algorithm for the identification of the aforementioned equivalence classes and the construction of the reduced ONV problem is outlined in Figure 3. As indicated in this figure, the construction of the reduced ONV problem, $\mathcal{E}^r = (X^r, x_0^r, x^0, X^{T^r}, \mathcal{A}^r, \mathcal{P}^r, \mathcal{N}^r)$, can be performed in three major stages: The first stage identifies and eliminates from the original stochastic graph $\mathcal{G} \equiv \mathcal{G}^0$, all those nodes that cannot be part of a closed communicating class that does not contain node x^0. The second stage processes the subgraph $\mathcal{G}^1$ returned by Stage 1, consisting of all nodes that can be isolated from node x^0 and their interconnecting actions, in order to partition it to the equivalence classes defined in the previous paragraph. This processing is done through an iterative computation which at every iteration selects one of the remaining nodes in the graph, computes the subgraph $\mathcal{G}'(x)$ corresponding to the equivalence class of x, and eliminates it from $\mathcal{G}^1$. The computation of $\mathcal{G}'(x)$ for any node x of $\mathcal{G}^1$ can be performed efficiently through concepts and techniques similar to those applied by the Ramadge & Wonham Supervisory Control theory [14] for

[5] possibly containing just x

establishing nonblocking behavior.[6] For the sake of brevity, we omit the details of this computation and we refer the interested reader to [15], [16]. The last stage of the algorithm presented in Figure 3 uses the results of the previous two stages in order to compile and return the stochastic graph $\mathcal{G}^r$ that represents the reduced ONV problem. An example execution of this algorithm is depicted in Figure 4.

We conclude this section with the statement of a property of the reduced ONV problem $\mathcal{E}^r$. This property is an immediate consequence of the construction of $\mathcal{G}^r$ and it will be very useful in the developments of the following sections.

Property 1: In the reduced ONV problem instance, $\mathcal{E}^r$, node x^0 is accessible from any target node $x \in X^{T^r}$, under any policy π^r defined on $\mathcal{G}^r$.

IV. A computationally efficient and asymptotically optimal policy for the reduced ONV problem

In this and the following section, we develop a series of computationally efficient and asymptotically optimal policies for the reduced ONV problem $\mathcal{E}^r$. However, in order to simplify the notation, we drop the index r from all the expressions presented in the subsequent discussion. We begin with the definition and study of a randomized policy that is obtained through a continuous – or *"fluid"* – relaxation of the MDP formulation corresponding to the reduced ONV problem. We shall refer to this policy as the policy π^{rel}. For reasons that will become clear in the subsequent discussion, it is convenient first to define and analyze policy π^{rel} for ONV problem instances with $x_0 = x^0$.

The "Relaxing LP" and the policy π^{rel} for reduced ONV problem instances with $x_0 = x^0$ The definition of the policy π^{rel} for a reduced ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with $x_0 = x^0$ relies on the optimal solution of the following LP formulation, that will be called the *"relaxing LP"*:

$$\min \sum_{a \in \mathcal{A}(x^0)} \chi_a \tag{5}$$

s.t.

$$\sum_{a:x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a = \sum_{a \in \mathcal{A}(x)} \chi_a, \ \forall x \in X \backslash \{x^0\} \tag{6}$$

$$\sum_{a:x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a \geq \mathcal{N}_x, \ \forall x \in X^T \tag{7}$$

$$\chi_a \geq 0, \ \forall a \in \bigcup_{x \in X} \mathcal{A}(x) \tag{8}$$

Equation 6 in the relaxing LP formulation enables a *"flow"*-based interpretation of its feasible solutions, $\{\chi_a\}$. This flow is defined on a graph that is obtained from the original graph $\mathcal{G}$, that characterizes the agent transitions among the nodes $x \in X$, by redirecting the transitions leading to node x^0, to a *"sink"* dummy node x^s. Fluid is pumped in this modified graph from node x^0, it is routed through the graph according to the flow pattern indicated by $\{\chi_a\}$, and eventually it gets absorbed to x^s through the redirected transitions. Under this flow-based interpretation, the objective of the formulation of Equations 5–8 is to minimize the amount of fluid pumped through x^0, while ensuring that the total flow entering each target node $x \in X^T$ is no less than the quantity expressed by the corresponding visitation requirement $\mathcal{N}_x$. Property 1 in Section III guarantees that this LP formulation is well-defined and it possesses a finite optimal value.

Given an optimal solution $\chi^* = \{\chi_a^* |\ a \in \bigcup_{x \in X} \mathcal{A}(x)\}$ of the LP defined by Equations 5-8, policy π^{rel} assigns to a state $s = (x, \mathcal{N}^c)$ with $\sum_{a \in \mathcal{A}(x)} \chi_a^* > 0$, an action $\pi^{rel}(x, \mathcal{N}^c) \in \mathcal{A}(x)$ according to the probability distribution

$$\text{Prob}(\pi^{rel}(x, \mathcal{N}^c) = a) = \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x)} \chi_a^*}, \ a \in \mathcal{A}(x) \tag{9}$$

On the other hand, states $s = (x, \mathcal{N}^c)$ with $\sum_{a \in \mathcal{A}(x)} \chi_a^* = 0$, are inaccessible under π^{rel}, and the policy is indeterminate at them. Clearly the deployment of the aforestated policy π^{rel} is of polynomial complexity with respect to the problem size $|\mathcal{E}|$.

The optimal value of the relaxing LP as a lower bound to V^* Let $e^{rel}_{x^0,j}$ denote the amount of flow reaching the target node $x^j \in X^T$ when a unit amount of flow is induced into the graph through node x^0 and it is conveyed according to the flow pattern defined by the routing probabilities of policy π^{rel} (cf. Eq. 9). Then, a simple conditioning argument can establish that $e^{rel}_{x^0,j}$ is equal to the expected number of visits to node x^j that take place when the control agent is initially placed at node x^0 and it is subsequently routed according to policy π^{rel} until it returns to node x^0. This interpretation of $e^{rel}_{x^0,j}$ leads to the following theorem:

Theorem 2: Consider a reduced ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with $x_0 = x^0$, and let V^*_{rel} and χ^* respectively denote the optimal value and an optimal solution of the relaxing LP. Also, let $e^{rel}_{x^0,j}$, $x^j \in X^T$, be defined on the basis of χ^* as indicated in the previous paragraph. Then,

$$V^*_{rel} = \max_{x^j \in X^T} \{\frac{\mathcal{N}_j}{e^{rel}_{x^0,j}}\} \leq V^* \tag{10}$$

The proof of this theorem is similar to the proof of Theorem 3 in [1], and it is omitted.

Establishing the asymptotic optimality of π^{rel} for reduced ONV problem instances with $x_0 = x^0$ Next we proceed to prove the asymptotic optimality of π^{rel} for reduced ONV problem instances with $x_0 = x^0$. For this, consider the problem sequence $\{\mathcal{E}(n)\}$ that is induced by a problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with

[6] The correspondence between these two problems can be seen more clearly when noticing that for the needs of the considered computation, a stochastic transition can be modeled by a transition to an intermediary dummy node with uncontrollable transitions to the possible outcomes. Then, in both cases, the problem reduces to confining the roaming agent to the maximal strongly connected component of the underlying graph that contains a certain node.

$x_0 = x^0$, through the scaling of the visitation requirement vector, $\mathcal{N}$, by a factor $n \in \mathbb{Z}^+$. Also, in the following, we shall let $\{V^*_{rel}(n)\}$ denote the sequence of the optimal objective values of the relaxing LPs implied by the problem sequence $\{\mathcal{E}(n)\}$, and $\{V^*(n)\}$ denote the sequence of the corresponding optimal expected total costs. Finally, we notice that the optimal solutions of the relaxing LP corresponding to problem instance $\{\mathcal{E}(n)\}$ are obtained by scaling the optimal solutions of the relaxing LP corresponding to the original problem instance $\{\mathcal{E}\}$ by a factor of n, which further implies the invariance of policy π^{rel} across the sequence $\{\mathcal{E}(n)\}$. Hence, we define $\{V^{\pi^{rel}}(n)\}$ as the sequence of the expected costs incurred by the application of the randomized policy π^{rel} to the problem instances $\mathcal{E}(n)$. Then, we have the following theorem:

Theorem 3: Given a reduced ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with $x_0 = x^0$, consider the problem sequence $\mathcal{E}(n)$ that is obtained through the uniform scaling of the visitation requirement vector $\mathcal{N}$ by a factor $n \in \mathbb{Z}^+$. Then, as $n \to \infty$,[7]

$$V^{\pi^{rel}}(n) - V^*_{rel}(n) = O(\sqrt{n}) \tag{11}$$

Furthermore, if there exists a target node x^k such that, for any other target node x^j,

$$\frac{\mathcal{N}_k}{e^{rel}_{x^0,k}} > \max_{j \neq k}\{\frac{\mathcal{N}_j}{e^{rel}_{x^0,j}}\} \tag{12}$$

then, as $n \to \infty$,

$$V^{\pi^{rel}}(n) - V^*_{rel}(n) = O(1) \tag{13}$$

Proof: In the subsequent discussion it is pertinent to decompose the motion of the control agent among the different nodes in X into a sequence of *"traversals"*, where the i^{th} traversal corresponds to the movements of the agent between the i^{th} and the $(i+1)^{st}$ visit by the agent to node x^0. Furthermore, we shall use the random variables Ξ^j_i, $i = 1, 2, \ldots$, to denote the random number of visits to the target node x^j during the i^{th} graph traversal under π^{rel}, and we shall let $\sigma^2_j = Var(\Xi^j_1)$, $j = 1, 2 \ldots, |X^T|$. Property 1 of the reduced ONV problem instances implies that σ^2_j is well-defined for all $j = 1, 2 \ldots, |X^T|$. Finally, $\{\psi^n_j,\ n \geq 0\}$ will denote a *renewal process* [6] associated with the sequence $\{\Xi^j_i : i = 1, 2, \ldots\}$, defined as

$$\psi^n_j = \max\{k : \sum_{i=1}^{k} \Xi^j_i \leq n \cdot \mathcal{N}_j\} \tag{14}$$

where $\psi^n_j = 0$ if $\Xi^j_1 > n \cdot \mathcal{N}_j$, $j : x^j \in X^T$. Then the performance of policy π^{rel} satisfies

$$V^{\pi^{rel}}(n) \leq E[\max_{j:x^j \in X^T}\{1 + \psi^n_j\}] \tag{15}$$

[7]We remind the reader that $f(n) = O(g(n)) \Rightarrow \exists c, n_0$ s.t. $0 \leq f(n) \leq c \cdot g(n),\ \forall n \geq n_0$.

Hence,

$$\begin{aligned} V^{\pi^{rel}}(n) - V^*_{rel}(n) &\leq \\ 1 + E[\max_{j:x^j \in X^T}\{\psi^n_j\}] - \max_{j:x^j \in X^T}\{\frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}\} &\leq \\ 1 + E[\max_{j:x^j \in X^T}\{|\psi^n_j - \frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}|\}] &\leq \\ 1 + \sum_{j:x^j \in X^T} E[|\psi^n_j - \frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}|] \end{aligned} \tag{16}$$

where the first inequality is the result of Equation 15 and Theorem 2, and the second inequality is the result of the following property:

$$\begin{aligned} \forall a_i, b_i \in R,\ i = 1, \ldots, n, \\ |\max\{a_1, a_2, \ldots, a_n\} - \max\{b_1, b_2, \ldots, b_n\}| \leq \\ \max\{|a_1 - b_1|, |a_2 - b_2|, \ldots, |a_n - b_n|\} \end{aligned} \tag{17}$$

From the *renewal central limit theorem* [6] we get that $\forall j : x^j \in X^T$,

$$\frac{1}{\sqrt{n}} \cdot (\psi^n_j - \frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}) \Rightarrow N(0, \frac{\sigma^2_j \cdot \mathcal{N}_j}{(e^{rel}_{x^0,j})^3}) \tag{18}$$

where '$\Rightarrow$' denotes convergence in distribution as $n \to \infty$ and $N(a, b)$ denotes the normal distribution with mean a and variance b. But then, Equation 18, when combined with Lemma 1 of the Appendix and the Continuous Mapping Theorem, imply that $\forall j : x^j \in X^T$,

$$\frac{1}{\sqrt{n}} E[|\psi^n_j - \frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}|] \longrightarrow E[|N(0, \frac{\sigma^2_j \cdot \mathcal{N}_j}{(e^{rel}_{x^0,j})^3})|] \tag{19}$$

as $n \to \infty$. Equation 11 now follows from Equation 16 when combined with Equation 19.

Next we prove Equation 13. For ease of presentation, assume that the node x^k of Equation 12 is the target node x^1. Then, we have that:

$$\begin{aligned} V^{\pi^{rel}}(n) - V^*_{rel}(n) &\leq \\ 1 + E[\max_{j:x^j \in X^T}\{\psi^n_j\}] - \max_{j:x^j \in X^T}\{\frac{n\mathcal{N}_j}{e^{rel}_{x^0,j}}\} &= \\ 1 + E[\max_{j:x^j \in X^T}\{\psi^n_j\}] - E[\psi^n_1] + E[\psi^n_1] - \frac{n\mathcal{N}_1}{e^{rel}_{x^0,1}} &= \\ 1 + E[\max_{j:x^j \in X^T}\{\psi^n_j - \psi^n_1\}] + E[\psi^n_1] - \frac{n\mathcal{N}_1}{e^{rel}_{x^0,1}} &\leq \\ 1 + \sum_{j \neq 1:x^j \in X^T} E[(\psi^n_j - \psi^n_1)^+] + E[\psi^n_1] - \frac{n\mathcal{N}_1}{e^{rel}_{x^0,1}} \end{aligned} \tag{20}$$

From Corollary 2.7.1 of [17], we get that

$$E[\psi^n_1] - \frac{n\mathcal{N}_1}{e^{rel}_{x^0,1}} \leq \frac{E[\Xi^2_1]}{2(e^{rel}_{x^0,1})^2} + \frac{1}{2e^{rel}_{x^0,1}} + o(1) \tag{21}$$

Next, we prove that $\forall j : x^j \in X^T$,

$$E[(\psi^n_j - \psi^n_1)^+] \to 0 \tag{22}$$

as $n \to \infty$. Indeed, for $r \geq 1$, $a_j^n = \frac{1}{\sqrt{n}}(\psi_j^n - \frac{n \cdot \mathcal{N}_j}{e_{x^0,j}^{rel}})$ and $c_j = \frac{\mathcal{N}_1}{e_{x^0,1}^{rel}} - \frac{\mathcal{N}_j}{e_{x^0,j}^{rel}} > 0$, we have that

$$\begin{aligned}
E[(\psi_j^n - \psi_1^n)^+] &= \\
E[(\psi_j^n - \psi_1^n) \cdot I(\psi_j^n \geq \psi_1^n)] &\leq \\
E[\psi_j^n \cdot I(\psi_j^n \geq \psi_1^n)] &\leq \\
\left[E[(\psi_j^n)^2] \cdot P(\psi_j^n \geq \psi_1^n)\right]^{1/2} &= \\
\Bigg[E[(\psi_j^n)^2] \cdot P((\psi_j^n - \frac{n \cdot \mathcal{N}_j}{e_{x^0,j}^{rel}}) - & \\
(\psi_1^n - \frac{n \cdot \mathcal{N}_1}{e_{x^0,1}^{rel}}) \geq \frac{n \cdot \mathcal{N}_1}{e_{x^0,1}^{rel}} - \frac{n \cdot \mathcal{N}_j}{e_{x^0,j}^{rel}})\Bigg]^{1/2} &= \\
\left[E[(\psi_j^n)^2] \cdot P(a_j^n - a_1^n \geq \sqrt{n} \cdot c_j)\right]^{1/2} &\leq \\
\left[E[(\psi_j^n)^2] \cdot \frac{1}{c_j^r \cdot n^{r/2}} \cdot E[(a_j^n - a_1^n)^r]\right]^{1/2} &\leq \\
\left[E[(\psi_j^n)^2] \cdot \frac{2^{r-1}}{c_j^r \cdot n^{r/2}} \cdot E[|a_j^n|^r + |a_1^n|^r]\right]^{1/2} & \qquad (23)
\end{aligned}$$

where the second inequality is an application of Schwarz inequality, the third inequality is an application of Markov inequality, and the last inequality is a direct consequence of $(a+b)^r \leq 2^{r-1} \cdot (|a|^r + |b|^r),\ a, b \in R$. Furthermore, from Theorem 2.3 of [17] we have that

$$E[(\psi_j^n)^2] = O(n^2) \qquad (24)$$

and if we choose r such that $\frac{r}{2} > 2$, then Equations 19, 23, 24 and Lemma 1 of the Appendix imply Equation 22.

Finally, Equation 13 follows immediately from Equation 20 when combined with Equations 21 and 22. □

The asymptotic optimality of policy π^{rel} for reduced ONV problem instances with $x_0 = x^0$ is an immediate implication of Theorem 3, a fact that is formally stated and proven in the following corollary:

Corollary 1: Given a reduced ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with $x_0 = x^0$, consider the problem sequence $\mathcal{E}(n)$ that is obtained through the uniform scaling of the visitation requirement vector $\mathcal{N}$ by a factor $n \in \mathbb{Z}^+$. Then, as $n \to \infty$,

$$\frac{V^{\pi^{rel}}(n)}{V^*(n)} \to 1 \qquad (25)$$

Proof: The combination of Theorems 2 and 3 implies that $\lim_{n\to\infty} \frac{V^{\pi^{rel}}(n)}{V^*(n)} \leq 1$, while the definition of V^* implies that $V^{\pi^{rel}}(n) \geq V^*(n)$, $\forall n \in \mathbb{Z}^+$. □

The policy π^{rel} for reduced ONV problem instances with $x_0 \neq x^0$ In the case of reduced ONV problem instances with $x_0 \neq x^0$, the relaxing LP is still defined by Equations 5–8. Furthermore, Equation 9 still provides an initial specification of the policy π^{rel}. However, since it is possible that the nitial state x_0 is among the states that are inaccessible under this specification of π^{rel}, it might be necessary to augment the policy specification with an action selection scheme over a subset of the states that are inaccessible under the original specification of the policy. In the case that such an augmentation of π^{rel} is necessary, the only requirements that are posed on it are (i) that it respects the original specification of π^{rel} over the nodes that this specification was initially defined, and (ii) that it renders node x^0 accessible from node x_0. Next, we state and prove the asymptotic optimality of the resulting policy.

Corollary 2: Given a reduced ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ with $x_0 \neq x^0$, consider the problem sequence $\mathcal{E}(n)$ that is obtained through the uniform scaling of the visitation requirement vector $\mathcal{N}$ by a factor $n \in \mathbb{Z}^+$. Then, as $n \to \infty$,

$$\frac{V^{\pi^{rel}}(n)}{V^*(n)} \to 1 \qquad (26)$$

Proof: Consider a reduced ONV problem instance $\mathcal{E}(n) = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, n\mathcal{N})$ with $x_0 \neq x^0$, and let $V^{\pi^{rel}}(n; x^0)$ and $V^*(n; x^0)$ respectively denote the values of the π^{rel} and the optimal policy when the process is started from x^0 instead of x_0. Then, we have:

$$\begin{aligned}
\frac{V^{\pi^{rel}}(n)}{V^*(n)} &\leq \frac{V^{\pi^{rel}}(n; x^0) + 1}{V^*(n)} \\
&= \frac{V^{\pi^{rel}}(n; x^0) + 1}{V^*(n; x^0)} \cdot \frac{V^*(n, x^0)}{V^*(n)} \qquad (27)
\end{aligned}$$

Corollary 1 implies that

$$\lim_{n\to\infty} \frac{V^{\pi^{rel}}(n; x^0) + 1}{V^*(n; x^0)} = 1 \qquad (28)$$

Similarly, Property 1 implies that

$$\lim_{n\to\infty} \frac{V^*(n, x^0)}{V^*(n)} = 1 \qquad (29)$$

But then, the result of Corollary 2 follows from Equations 27–29, when noticing that $\frac{V^{\pi^{rel}}(n)}{V^*(n)} \geq 1$. □

We conclude this section by noticing that the results of Theorem 3 and their derivation imply that, under the condition that there exists a k such that $\frac{\mathcal{N}_k}{e_{x^0,k}^{rel}} > \max_{j \neq k}\{\frac{\mathcal{N}_j}{e_{x_0,j}^{rel}}\}$, the performance of π^{rel} and π^* will differ from the lower bound $V^{rel}(n)$ by at most a constant K, as the scaling factor n grows to infinity. A similar result can be established for the case of reduced ONV problem instances with $x_0 \neq x^0$, through a conditioning argument similar to that used in the proof of Corollary 2. An intuitive interpretation of these two results can be obtained by considering the ratio $\frac{\mathcal{N}_j}{e_{x^0,j}^{rel}}$ to be a "measure of difficulty" of the visitation requirement of the target node x^j, in the corresponding problem instance with $x_0 = x^0$. As n grows to infinity, the differences $\frac{n \cdot \mathcal{N}_k}{e_{x^0,k}^{rel}} - \frac{n \cdot \mathcal{N}_j}{e_{x^0,j}^{rel}}$ are also growing, hence the solution of the relaxing LP contains enough information in order to bias the system behavior towards

the optimal solution. On the other hand, when the target nodes corresponding to the maximal ratio $\frac{n \cdot \mathcal{N}_k}{e^{rel}_{x^0,k}}$ are more than one, π^{rel} will treat those nodes as equally difficult targets. Furthermore, the *static* nature of this policy will not allow it to exploit the dynamics of the future problem states, where the original ties will have been resolved. This last observation motivates the study of *adaptive* implementations of π^{rel}, where the routing probabilities that define the new policy are revised at every change of the vector $\mathcal{N}^c$. The possibilities and the challenges for the development of such adaptive control schemes for the ONV problem are addressed in the following section.

V. Adaptive Policies

In this section we consider briefly the possibilities and challenges for enhancing the performance of policy π^{rel} through more *dynamic* implementations that adjust the policy logic in real-time, on the basis of the available information regarding the evolution of the system state. A first possibility for such an improvement upon π^{rel} is the implementation of a *"rollout"* scheme that uses π^{rel} as its *"base"* policy [2]. More specifically, a rollout implementation of π^{rel} essentially clusters the states of the underlying MDP with the same vector $\mathcal{N}^c$ in *"macro-states"*,[8] and every time that a new macro-state is entered, it computes a locally optimized policy for that macro-state. The computation of this localized policy for any visited macro-state is based upon the restricted application of some standard dynamic programming approach on the subspace spanned by the macro-state, combined with the further assumption that policy π^{rel} will be followed outside that region (and therefore, the values of the states s that can be reached from the considered macro-state through a single transition are taken to be equal to $V^{\pi^{rel}}(s)$). Using arguments similar to those provided in [2], it is easy to show that such a scheme will result in improved performance compared to the performance attained by the original implementation of π^{rel}. On the other hand, a potential source of difficulty for this rollout-based implementation of π^{rel} in the considered ONV problem context stems from the fact that the aforementioned $V^{\pi^{rel}}(s)$ values, that are used for the policy specification at the visited macro-states, must be obtained through simulation.[9]

An alternative dynamic implementation of π^{rel} for the considered problem contexts will seek to revise the routing probabilities every time a visitation requirement is satisfied, by formulating and re-solving the relaxing LP corresponding to that particular state. We shall refer to this *adaptive* implementation of π^{rel} as π^{adrel}. From a computational standpoint, π^{adrel} is definitely a much more tractable proposition than the rollout-based implementation of π^{rel}. Furthermore, some computational studies reported in the next section indicate that π^{adrel} presents excellent performance, typically outperforming any other suboptimal policy applied on that problem. On the other hand, our theoretical understanding of the dynamics underlying this policy is currently limited; in particular, currently we lack any theoretical guarantee that $V^{\pi^{adrel}} \leq V^{\pi^{rel}}$. The thorough analysis of the dynamics of policy π^{adrel} is an interesting and challenging task, and it is part of our current investigations.

[8] Figure 2 provides an example of such a state space partitioning to "macro-states"

[9] A special case where these $V^{\pi^{rel}}(s)$ values can be obtained through closed formulae, is discussed in the next section.

VI. Closed-form performance evaluation of π^{rel} for specially structured ONV problem instances and its implications

In this section we focus on the more restricted ONV problem version that was initially studied in [1], and we show that in this case, the expected performance of π^{rel} admits a closed-form characterization. Furthermore, we discuss the practical possibilities offered by this result for an enhanced implementation of π^{rel} in the considered problem context.

The ONV problem considered in the following is obtained from the original definition of the ONV problem presented in Section II, through the addition of the following assumptions:

Assumption 1: $x_0 = x^0$.

Assumption 2: The transitions out from the target nodes $x^j \in X^T$ lead deterministically to node x^0.

Assumption 3: The stochastic subgraph $\mathcal{G}'$, that is obtained by the elimination of the transitions emanating from all $x^j \in X^T$, has an *acyclic* structure, with node x^0 defining its unique *"root"* node and the target nodes $x^j \in X^T$ being a subset of the *"terminal"* nodes.

It is easy to see that the ONV problem instances satisfying Assumptions 1–3 are in reduced form. Also, the traversal of graph $\mathcal{G}'$, between two consecutive visits to the node x^0, can satisfy at most one visitation requirement with respect to a single node. Therefore, the parameters $e^{rel}_{x^0,j}$ $x^j \in X^T$, essentially express the probability that target node x^j will be visited during a single traversal of $\mathcal{G}'$ under policy π^{rel}. The next theorem builds upon these observations in order to provide a closed-form evaluation of π^{rel}.

Theorem 4: Consider the implementation of policy π^{rel} on an ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ satisfying Assumptions 1–3. Then we have:

$$V^{\pi^{rel}} = E[\max_{j:\mathcal{N}_j>0} \{\frac{1}{e^{rel}_{x^0,j}} \sum_{i=1}^{\mathcal{N}_j} \Xi^i_j\}] \tag{30}$$

where Ξ^i_j are independent identically distributed exponential random variables with rate $\lambda = 1$.

Proof: Consider a continuous-time variation of the problem where the graph traversals are guided by policy π^{rel}, and their durations, Y_i, are generated by a Poisson process with rate $\lambda = 1$. Then, the visits to each target node $x^j \in X^T$ define a Poisson process with rate $e^{rel}_{x^0,j}$, and these Poisson processes are independent [6]. Let T_j denote the time until target node x^j has satisfied its visitation require-

ments, and N denote the total number of graph traversals required until every visitation requirement is satisfied. Then, (i) each T_j is distributed according to a Gamma distribution with parameters $\mathcal{N}_j$ and $e^{rel}_{x^0,j}$, and (ii) the T_j's are independent [6]. Set $T = \max_{j:\mathcal{N}_j>0}\{T_j\}$. Then

$$\begin{aligned} E[\max_{j:\mathcal{N}_j>0}\{T_j\}] &= E[T] \\ &= E[\sum_{i=1}^{N} Y_i] \\ &= E[E[\sum_{i=1}^{N} Y_i|N]] \\ &= E[N \cdot E[Y_1]] \\ &= E[N] \\ &= V^{\pi^{rel}} \end{aligned} \tag{31}$$

Since each T_j is equal in distribution to $\frac{1}{e^{rel}_{x^0,j}}\sum_{i=1}^{\mathcal{N}_j}\Xi^i_j$, we have that

$$E[\max_{j:\mathcal{N}_j>0}\{T_j\}] = E[\max_{j:\mathcal{N}_j>0}\{\frac{1}{e^{rel}_{x^0,j}}\sum_{i=1}^{\mathcal{N}_j}\Xi^i_j\}] \tag{32}$$

The result now follows by combining Equations 31 and 32. □

Theorem 4 further implies that $V^{\pi^{rel}}$ can be obtained through the numerical integration of a continuous function since

$$\begin{aligned} V^{\pi^{rel}} &= E[\max_{j:\mathcal{N}_j>0}\{\frac{1}{e^{rel}_{x^0,j}}\sum_{i=1}^{\mathcal{N}_j}\Xi^i_j\}] \\ &= \int_0^\infty P(\max_{j:\mathcal{N}_j>0}\{\frac{1}{e^{rel}_{x^0,j}}\sum_{i=1}^{\mathcal{N}_j}\Xi^i_j\} > t)dt \\ &= \int_0^\infty (1 - \prod_{j:\mathcal{N}_j>0} P(\frac{1}{e^{rel}_{x^0,j}}\sum_{i=1}^{\mathcal{N}_j}\Xi^i_j \le t))dt \\ &= \int_0^\infty (1 - \prod_{j:\mathcal{N}_j>0} F_{\mathcal{N}_j}(e^{rel}_{x^0,j} \cdot t))dt \end{aligned} \tag{33}$$

where $F_{\mathcal{N}_j}(t)$ is the cumulative distribution function of the Gamma$(\mathcal{N}_j, 1)$ distribution. Equation 33 can be especially useful in a rollout implementation of π^{rel} along the lines suggested in Section V.

Furthermore, it is easy to see that the above analysis applies to any other randomized policy π that, similar to π^{rel}, (i) bases the action selection probabilities at every state $s = (x, \mathcal{N}^c)$ only upon the component x of s,[10] and (ii) maintains a positive probability, $e^{\pi}_{x^0,j}$, for reaching each target node $x^j \in X^T$ during a single traversal of $\mathcal{G}'$. It can be easily shown [18] that the space Π^S of these static randomized policies is in one-to-one correspondence with the space of vectors $\mathcal{X} = \{\chi_a |\ a \in \mathcal{A}(x), x \in X\backslash X^T\}$ satisfying

$$\sum_{a\in\mathcal{A}(x^0)} \chi_a = 1 \tag{34}$$

$$\forall x \in X\backslash\{x^0, X^T\},$$

$$\sum_{a:x\in\mathcal{S}(a)} \chi_a \cdot p(x,a) = \sum_{a\in\mathcal{A}(x)} \chi_a \tag{35}$$

$$\forall x \in X^T \text{ with } \mathcal{N}_x > 0,$$

$$\sum_{a:x\in\mathcal{S}(a)} \chi_a \cdot p(x,a) > 0 \tag{36}$$

The variables $\chi_a,\ a \in \mathcal{A}(x),\ x \in X\backslash X^T$, that appear in the above formulation, denote the probability of executing action a during any single traversal of graph $\mathcal{G}'$ under the corresponding policy π, and therefore,

$$e^{\pi}_{x^0,j} = \sum_{a:x^j\in\mathcal{S}(a)} \chi^{\pi}_a \cdot p(x^j, a),\ x^j \in X^T \tag{37}$$

Hence, the optimization problem $\min_{\pi\in\Pi^S} V^{\pi}$ can be expressed as

$$\begin{aligned} &\min V(e^{\pi}_{x^0,j}) \\ &\text{s.t.} \quad e^{\pi}_{x^0,j} = e^{\pi}_{x^0,j}(\chi_a), \quad \chi_a \in \mathcal{X} \end{aligned} \tag{38}$$

Equations 34–37 imply that the solution space of this last problem is convex, while the convexity of its objective function is implied by Equation 33. Hence, the optimization problem defined by Equation 38 possesses a convex smooth structure and therefore it can be effectively addressed by standard solution techniques coming from the area of nonlinear programming; we refer to [19] for the relevant details. An optimal solution for the formulation of Equation 38 will be denoted by χ^{opt}, and the corresponding randomized policy by π^{opt}. Clearly, $V^{\pi^{opt}} \le V^{\pi^{rel}}$.

As a last result we show that the definition of π^{opt} through the formulation of Equation 38 enables also an effective characterization of the gains achieved through its adaptive implementation according to an adaptation mechanism similar to that applied by policy π^{adrel}; i.e., under this adaptive implementation of π^{opt}, the action selection probabilities are re-computed, by resolving the formulation of Equation 38, every time that another visitation requirement is satisfied, and the underlying process enters a new macro-state. We shall refer to the resulting randomized policy as π^{adopt}. For π^{adopt}, we have the following theorem:[11]

Theorem 5: For any ONV problem instance $\mathcal{E} = (X, x_0, x^0, X^T, \mathcal{A}, \mathcal{P}, \mathcal{N})$ satisfying Assumptions 1–3, $V^{\pi^{adopt}} \le V^{\pi^{opt}}$

Proof: We prove this result by induction on $|\mathcal{N}|$, i.e., the total number of visitation requirements. For $|\mathcal{N}| = 1$, the process will visit only one macro-state before its termination, and therefore, $V^{\pi^{adopt}} = V^{\pi^{opt}}$. Next, we

[10] We remind the reader that we have referred to these policies as *static*

[11] We remind the reader that, as remarked in Section V, currently we lack a similar result for policy π^{adrel}.

assume that the inequality of Theorem 5 holds for $|\mathcal{N}| \leq n$, and we show that it will also hold for $|\mathcal{N}| = n+1$. To obtain this result, we first notice that the value function of any proper policy π will satisfy the following recursion:

$$V^{\pi}(x^0,\mathcal{N}) = \frac{1}{\sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi(\mathcal{N})}_{x^0,j}} \cdot [1 + \sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi(\mathcal{N})}_{x^0,j} \cdot V^{\pi}(x^0,\mathcal{N}-\mathbf{1}_{x^j})] \quad (39)$$

where (i) $e^{\pi(\mathcal{N})}_{x^0,j}$ denotes the probability of reaching node $x \in X^T$ in any single traversal of graph $\mathcal{G}'$ under policy π, while starting from state $(x^0,\mathcal{N})$ (cf. Equation 37), and (ii) $\mathbf{1}_{x^j}$ denotes the unit vector of dimensionality equal to $|X^T|$ and with its non-zero component corresponding to node x^j. Application of Equation 39 to π^{adopt} gives that

$$V^{\pi^{adopt}(\mathcal{N})}(x^0,\mathcal{N}) = \frac{1}{\sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi^{adopt}(\mathcal{N})}_{x^0,j}} \cdot [1 + \sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi^{adopt}(\mathcal{N})}_{x^0,j} \cdot V^{\pi^{adopt}(\mathcal{N})}(x^0,\mathcal{N}-\mathbf{1}_{x^j})] \quad (40)$$

However, the definition of π^{adopt} implies that $e^{\pi^{adopt}(\mathcal{N})}_{x^0,j} = e^{\pi^{opt}(\mathcal{N})}_{x^0,j}$ and $V^{\pi^{adopt}(\mathcal{N})}(x^0,\mathcal{N}-\mathbf{1}_{x^j}) = V^{\pi^{adopt}(\mathcal{N}-\mathbf{1}_{x^j})}(x^0,\mathcal{N}-\mathbf{1}_{x^j})$, for all $x^j \in X^T$. Furthermore, $V^{\pi^{adopt}(\mathcal{N}-\mathbf{1}_{x^j})}(x^0,\mathcal{N}-\mathbf{1}_{x^j}) \leq V^{\pi^{opt}(\mathcal{N}-\mathbf{1}_{x^j})}(x^0,\mathcal{N}-\mathbf{1}_{x^j}) \leq V^{\pi^{opt}(\mathcal{N})}(x^0,\mathcal{N}-\mathbf{1}_{x^j})$, $\forall x^j \in X^T : \mathcal{N}_j > 0$, where the first inequality results from the induction hypothesis and the second from the definition of π^{opt}. But then, Equation 40 implies that

$$V^{\pi^{adopt}(\mathcal{N})}(x^0,\mathcal{N}) \leq \frac{1}{\sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi^{opt}(\mathcal{N})}_{x^0,j}} \cdot [1 + \sum_{x^j \in X^T:\mathcal{N}_j>0} e^{\pi^{opt}(\mathcal{N})}_{x^0,j} \cdot V^{\pi^{opt}(\mathcal{N})}(x^0,\mathcal{N}-\mathbf{1}_{x^j})] = V^{\pi^{opt}(\mathcal{N})}(x^0,\mathcal{N}) \quad (41)$$

□

Example We conclude this section by reporting a computational study on two ONV problem instances that exemplifies and highlights the results presented in this paper. The two considered problem instances satisfy the Assumptions 1–3 stated at the beginning of this section, and they are defined by the acyclic stochastic graph $\mathcal{G}'$ depicted in Figure 5 and the respective visitation requirement vectors $\mathcal{N} = (3,1,1,0,0)$ and $\mathcal{N} = (1,2,2,2,1)$, which associate a visitation requirement to each of the terminal nodes, x^j, $j = 4,\ldots,8$. The solution of the corresponding relaxing LPs indicates that the problem instance defined by $\mathcal{N} = (3,1,1,0,0)$ satisfies the condition of Equation 12 in Theorem 3, with the most difficult visitation requirement determined by the terminal node x^4. On the other hand, the problem instance defined by $\mathcal{N} = (1,2,2,2,1)$ has a constant ratio $\mathcal{N}_j/e^{rel}_{x^0,j}$ across all $j = 4,\ldots,8$. Figures 6 and 7 report the performance of the policies π^{rel},

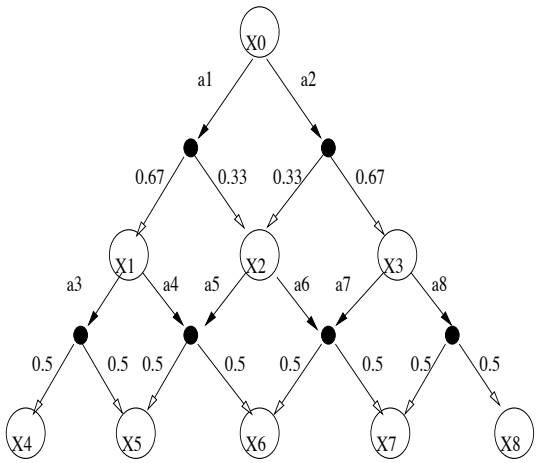

Fig. 5. Exampe – The stochastic graph for the considered problem instances

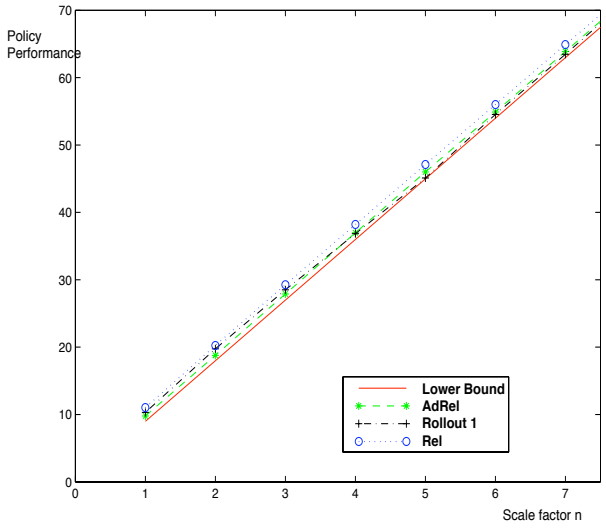

Fig. 6. Example – The performance of various simple and adaptive randomized policies compared to the lower bound $V^*_{rel}(n)$, for the basic visitation requirement vector $\mathcal{N} = (3,1,1,0,0)$ and $n = 1,\ldots,7$

π^{adrel} and π^{roll} in each of these two cases, as the corresponding vector $\mathcal{N}$ is scaled to increasingly larger values. The reported values for the policy π^{rel} were obtained from Equation 33. The performance of the policies π^{adrel} and π^{roll} was estimated through simulation. As expected from Theorem 3, in the case of the visitation requirement vector $\mathcal{N} = (3,1,1,0,0)$, the performance of all three policies converges very fast to the lower bound $V^*_{rel}(n)$ – cf. Figure 6.[12] On the other hand, the ties of the ratios $\mathcal{N}_j/e^{rel}_{x^0,j}$, $j = 4,\ldots,8$, in the case of the visitation requirement vector $\mathcal{N} = (1,2,2,2,1)$, result in the divergence of the performance of the considered policies from the lower bound $V^*_{rel}(n)$ – cf. Figure 7. However, as expected, the distance of the performance of these policies from $V^*_{rel}(n)$ increases in a slow, sub-linear manner with respect to n, so that the corresponding ratios $V^{\pi}(n)/V^*_{rel}(n)$ decrease to one. Fi-

[12] A closer examination of the proof of Theorem 3 will reveal that for ONV problem instances satisfying Assumptions 1–3 and the condition of Equation 12, $V^{\pi^{rel}}(n) - V^*_{rel}(n) \to 0$, as $n \to \infty$.

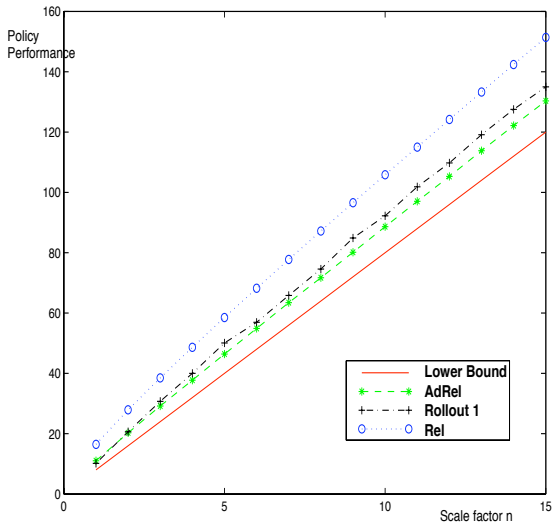

Fig. 7. Example – The performance of various simple and adaptive randomized policies compared to the lower bound $V^*_{rel}(n)$, for the basic visitation requirement vector $\mathcal{N} = (1, 2, 2, 2, 1)$ and $n = 1, \ldots, 15$

nally, it is worth-noticing that π^{adrel} outperforms the other two policies, demonstrating a performance that is pretty close to the lower bound $V^*_{rel}(n)$.

VII. Conclusions

This paper revisited the problem of optimal node visitation in stochastic digraphs, that was originally introduced in [1], and it extended the relevant theory (i) by addressing a much broader set of problem structures, and (ii) by enriching and improving the class of asymptotically optimal policies available for it. In addition, a series of novel and/or stronger properties for the performance of these policies were derived. The presented results are motivated by and are similar in spirit to some recent developments in stochastic scheduling theory and the suboptimal control of Markov Decision Processes. Future work will seek to (i) formally analyze the computational complexity of the considered problem; (ii) capitalize upon further insights and results from stochastic scheduling theory, like those presented in [20], in order to identify additional structure and properties for it; and (iii) extend the results and the policies developed herein to other problem variations, like in the case that each graph traversal might generate more than one threads executing in parallel on the underlying stochastic graph $\mathcal{G}$ that defines the problem.

Appendix

Lemma 1: Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $E[X_1^r]$ exists for every $r \geq 1$ and $\mu = E[X_1]$. Set $S_0 = 0$, $S_k = \sum_{i=1}^{k} X_i$ and define $\psi_n = \max\{k : S_k \leq n \cdot c\}$. Then

$$\{n^{-r/2}(\psi_n - \frac{n \cdot c}{\mu})^r, n \geq 1\} \tag{42}$$

is uniformly integrable for every $r \geq 1$.

Proof: Let $\psi'_n = \min\{k : S_k > n \cdot c\}$. Then ψ'_n is a stopping time and, from Lemma 2.3 of [17], we have that

$$E[(\sum_{i=1}^{\psi'_n}(X_i - \mu))^r] \quad \leq \quad C(r, E[X^r]) \cdot E[(\psi'_n)^{r/2}] \tag{43}$$

where $C(r, E[X^r])$ is a constant depending only on r and $E[X^r]$. Equation 43 further implies that

$$E[n^{-r/2} \cdot (\sum_{i=1}^{\psi'_n}(X_i - \mu))^r] \leq$$
$$C(r, E[X^r]) \cdot E[(\frac{\psi'_n}{n})^{r/2}] \tag{44}$$

From Equation 44 and Theorem 2.3 of [17], we get

$$\sup_{n \geq 1} E[n^{-r/2} \cdot (\sum_{i=1}^{\psi'_n}(X_i - \mu))^r] < \infty \tag{45}$$

which implies the uniform integrability of $\{n^{-r/2} \cdot (\sum_{i=1}^{\psi'_n}(X_i - \mu))^r, n \geq 1\}$ [21].

Next, consider the quantity $S_{\psi'_n} - n \cdot c$, i.e., the excess over the boundary for the renewal process ψ_n, and notice that for every $r \geq 1$,

$$|S_{\psi'_n} - n \cdot c|^r \leq |X_{\psi'_n}|^r \leq \sum_{i=1}^{\psi'_n} |X_i|^r \tag{46}$$

When combined with Wald's equation [6], Equation 46 further implies that

$$E[|S_{\psi'_n} - n \cdot c|^r] \leq E[\psi'_n] E[|X_1|^r] \tag{47}$$

and for $r > 2$, we have that

$$\frac{E[|S_{\psi'_n} - n \cdot c|^r]}{n^{r/2}} \leq \frac{E[\psi'_n]}{n} \frac{E[|X_1|^r]}{n^{r/2-1}} \tag{48}$$

Since from Corollary 2.3.1 of [17] we know that $\sup_n E[\psi'_n]/n < \infty$, we can also claim that

$$\sup_n \frac{E[|S_{\psi'_n} - n \cdot c|^{r+\epsilon}]}{n^{(r+\epsilon)/2}} < \infty \tag{49}$$

and therefore, the sequence $\{n^{-r/2} \cdot |S_{\psi'_n} - n \cdot c|^r, n \geq 1\}$ is uniformly integrable.

By the definition of the renewal process ψ'_n,

$$n^{-1/2} \cdot \sum_{i=1}^{\psi'_n}(X_i - \mu) \leq$$
$$n^{-1/2} \cdot (n \cdot c - \mu \cdot \psi'_n) \leq$$
$$n^{-1/2} \cdot \sum_{i=1}^{\psi'_n}(X_i - \mu) + n^{-1/2} \cdot (S_{\psi'_n} - n \cdot c) \tag{50}$$

which further implies that

$$|n^{-1/2} \cdot (n \cdot c - \mu \cdot \psi'_n)| \leq$$
$$|n^{-1/2} \cdot \sum_{i=1}^{\psi'_n}(X_i - \mu)| + |n^{-1/2} \cdot (S_{\psi'_n} - n \cdot c)| \tag{51}$$

Combined with the inequality $(a+b)^r \leq 2^{r-1}\cdot(|a|^r + |b|^r),\ a,b \in R$, Equation 51 gives

$$|n^{-1/2}(n\cdot c - \mu\cdot\psi'_n)|^r \leq$$
$$2^{r-1}\cdot(|n^{-1/2}\sum_{i=1}^{\psi'_n}(X_i-\mu)|^r + |n^{-1/2}\cdot(S_{\psi'_n}-n\cdot c)|^r) \quad (52)$$

But then, the uniform integrability of $\{n^{-r/2}\cdot(\sum_{i=1}^{\psi'_n}(X_i-\mu))^r, n\geq 1\}$ and $\{n^{-r/2}\cdot|S_{\psi'_n}-n\cdot c|^r, n\geq 1\}$ and Equation 52 imply the uniform integrability of $\{n^{-r/2}\cdot(n\cdot c-\mu\cdot\psi'_n)^r, n\geq 1\}$. Since $\psi'_n = \psi_n + 1$ we have that,

$$n^{-1/2}\cdot(n\cdot c-\mu\cdot\psi_n) =$$
$$n^{-1/2}\cdot(n\cdot c-\mu\cdot\psi'_n) + n^{-1/2}\cdot\mu \quad (53)$$

which gives

$$n^{-r/2}\cdot|n\cdot c-\mu\cdot\psi_n|^r \leq$$
$$2^{r-1}\cdot(n^{-r/2}\cdot|n\cdot c-\mu\cdot\psi'_n|^r + n^{-r/2}\cdot\mu^r) \quad (54)$$

and implies the uniform integrability of $\{n^{-r/2}\cdot(n\cdot c-\mu\cdot\psi_n)^r,\ n\geq 1\}$. □

References

[1] T. Bountourelis and S. Reveliotis, "Optimal node visitation in acyclic stochastic digraphs," in *Proceedings the 8th Intl Workshop on Discete Event Systems (WODES'06)*. IFAC, 2006, pp. 358–365.

[2] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

[3] D. Bertsimas and D. Gamarnik, "Asymptotically optimal algorithms for job shop scheduling and packet switching," *Journal of Algorithms*, vol. 33, pp. 296–318, 1999.

[4] D. Bertsimas and J. Sethuraman, "From fluid relaxations to practical algorithms for job shop scheduling: The makespan objective," *Mathematical Programming*, vol. 92, pp. 61–102, 2002.

[5] J. Niño–Mora, "Stochastic scheduling," in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Kluwer, 2001, pp. 367–372.

[6] S. M. Ross, *Stochastic Processes*. NY: Wiley and Sons, 1996.

[7] P. Glasserman and D. Yao, *Monotone Structure in Discrete-Event Systems*. NY,NY: John Wiley & Sons, Inc., 1994.

[8] D. P. Bertsekas, "Dynamic programming and suboptimal control: A survey from ADP to MPC," *European Journal of Control*, vol. 11, pp. 310–334, 2005.

[9] M. Pinedo, *Scheduling: Theory, Algorithms and Systems (2nd ed.)*. Upper Saddle River, NJ: Prentice Hall, 2002.

[10] J. G. Dai, "Stability of fluid and stochastic processing networks,," Center for Mathematical Physics and Stochastics, University of Aarhus, Denmark, Tech. Rep. ISSN 1398-7957, 1999.

[11] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. NY,NY: Sprigner, 2001.

[12] S. Meyn, *Control Techniques for Complex Networks*. Cambridge, UK: Cambridge University Press, 2008.

[13] S. A. Reveliotis and T. Bountourelis, "Efficient PAC learning for episodic tasks with acyclic state spaces," *Journal of Discrete Event Systems: Theory and Applications*, vol. 17, pp. 307–327, 2007.

[14] P. J. G. Ramadge and W. M. Wonham, "The control of discrete event systems," *Proceedings of the IEEE*, vol. 77, pp. 81–98, 1989.

[15] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*. Boston, MA: Klumwer Academic Pub., 1999.

[16] S. A. Reveliotis, *Real-time Management of Resource Allocation Systems: A Discrete Event Systems Approach*. NY, NY: Springer, 2005.

[17] A. Gut, "On the moments and limit distibutions of some first passage times," *The Annals of Probability*, vol. 2, No. 2, pp. 277–308, 1974.

[18] T. Bountourelis and S. Reveliotis, "Rollout policies for the problem of optimal node visitation in acyclic stochastic digraphs," in *European Control Conference 2007*. IEEE, 2007, pp. 2456–2463.

[19] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA.: Athena Scientific, 1995.

[20] D. Bertsimas and J. Niño–Mora, "Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems," *Mathematics of Operations Research*, vol. 21, pp. 257–306, 1996.

[21] P. Billingsley, *Convergence of probability measures*. NY: Wiley and Sons, 1968.