Optimal node visitation in acyclic stochastic digraphs

Theologos Bountourelis and Spyros Reveliotis
School of Industrial & Systems Engineering
Georgia Institute of Technology
{tbountou,spyros}@isye.gatech.edu

Abstract

Given a stochastic, acyclic, connected digraph with a single source node and a control agent that repetitively traverses this graph, each time starting from the source node, we want to define a control policy that will enable this agent to visit each of the graph terminal nodes a prespecified number of times, while minimizing the expected number of the graph traversals. We first formulate this problem as a specially structured Discrete Time Markov Decision Process, and subsequently we develop an asymptotically optimal randomized policy of polynomial complexity with respect to the problem size. Finally, two further outcomes of this analysis are a lower and an upper bound to the value of the optimal policy, V^*.

1 Introduction

The problem addressed in this work can be stated as follows: Given a stochastic, acyclic, connected digraph with a single source node and a control agent that repetitively traverses this graph, each time starting from the source node, we want to define a control policy that will enable this agent to visit each of the graph terminal nodes a prespecified number of times, while minimizing the expected number of the graph traversals. From a practical standpoint, this problem arises, for instance, in 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 its various stages. Under the assumption 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. A similar problem also arises while trying to learn an optimal policy for a sequential decision making process over a stochastic acyclic state space. In that case, the learning agent tries to obtain a series of observations of the values of the various decisions made at the different

problem states, while minimizing the number of the executed process runs; we refer the reader to [5] for further details on this application.

In this work first we provide a detailed formulation of the aforestated problem as a specially structured Markov Decision Process (MDP) [1]. However, because the solution of this MDP through standard techniques provided by the MDP theory is of non-polynomial complexity with respect to the size of the problem-defining elements, we also develop a randomized policy that can be derived and implemented in polynomial time, and it is asymptotically optimal; more specifically, the ratio of the value of this policy to the value of the optimal policy converges to unity, as the non-zero node visitation requirements grow uniformly to infinity. Finally, a last contribution of the presented work, that results as a by-product of the aforementioned developments, is the establishment of a lower and an upper bound for the value of the optimal policy.

From a presentational standpoint, this material is organized as follows: Section 2 provides a formal characterization of the problem considered in this work, and proceeds to its formulation and solution as a *"stochastic shortest path (SSP)"* problem [1]. It also points out the limitations arising from the non-polynomial complexity of the standard SSP solution approach and the possibility of alleviating this computational complexity by taking advantage of some underlying special structure. Section 3 introduces the suboptimal but computationally efficient policy mentioned in the earlier paragraph, and proves its asymptotic optimality. In the process, it also derives a lower bound to the value of the optimal policy. Deriving an upper bound to this value is the topic of Section 4. Finally, Section 5 concludes the paper and suggests directions for future work.

2 Problem description and its MDP formulation

This section first provides a formal characterization of the defining elements of the considered problem and subsequently it proceeds to its rigorous formulation and solution, based upon concepts and techniques borrowed from the MDP theory [1]. The last part of this section also considers the computational complexity of the presented solution approach and motivates the need for the suboptimal but computationally more efficient solution approach developed in Section 3.

The defining problem elements An instance of the problem considered in this work is completely defined by a quadruple $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$, where

- X is a finite set of *nodes*, that is partitioned into a sequence of *"layers"*, X^0, X^1, ...,

X^L. $X^0 = \{x^0\}$ defines the *source* or *root node*, while nodes $x \in X^L$ are the *terminal* or *leaf* nodes.

- $\mathcal{A}$ 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, $\mathcal{S}(a)$, of the distributions $p(\cdot; a)$ are subsets of the set X that satisfy the following property: For any given action $a \in \mathcal{A}(x)$ with $x \in X^i$ for some $i = 0, \ldots, L-1$, $\mathcal{S}(a) \subseteq \bigcup_{j=i+1}^{L} X^j$; for $a \in \mathcal{A}(x)$ with $x \in X^L$, $\mathcal{S}(a) = X^0$. In words, the previous assumption implies that the control agent traverses the space defined by the node set X in an iterative manner, where each iteration is an *"acyclic"* traversal; more specifically, the sequence of nodes visited during each such iteration starts from the root node, x^0, ends at a leaf node, $x \in X_L$, and it is monotonically increasing with respect to the layer of the intermediately visited nodes. Furthermore, it is assumed that for every node $x \in X$, there exists at least one action sequence $\xi(x) = a^{(0)} a^{(1)} \ldots a^{(k(x))}$ such that (i) $a^{(0)} \in A(x^0)$, (ii) $\forall i = 1, \ldots, k(x)$, $a^{(i)} \in A(x^{(i)})$ with $p(x^{(i)}; a^{(i-1)}) > 0$, and (iii) $p(x; a^{k(x)}) > 0$; we shall refer to this action sequence as an *action path* from node x^0 to node x.

- $\mathcal{N}$ is the *visitation requirement vector*, that associates with each node $x \in X^L$ a visitation requirement $\mathcal{N}_x \in \mathbb{Z}_+ \cup \{0\}$. The *support* $||\mathcal{N}||$ of $\mathcal{N}$ is defined by the nodes $x \in X^L$ with $\mathcal{N}_x > 0$; we shall refer to nodes $x \in ||\mathcal{N}||$ as the problem *"target"* nodes.

- Finally, 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.

For the purposes of the subsequent discussion it is pertinent to perceive the node space X endowed with the transition function $\mathcal{P}$ as an *"acyclic stochastic digraph"*, $\mathcal{G}$, where the node set of $\mathcal{G}$ is defined by X and its arcs are defined by the restriction of $\mathcal{P}$ on $\bigcup_{x \in X \setminus X^L} \mathcal{A}(x)$. We shall also employ the variable vector $\mathcal{N}^c$ to denote the *vector of the remaining visitation requirements*. The control agent starts from the initial 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, on the graph, and the vector of the remaining node visitation requirements, $\mathcal{N}^c$, (ii) selects an action $a \in \mathcal{A}(x)$ and commands its execution, and (iii) upon reaching one of the terminal nodes, $x \in X^L$, updates $\mathcal{N}_x{}^c$ to $(\mathcal{N}_x{}^c - 1)^+$, and subsequently, *resets* itself back to the initial node

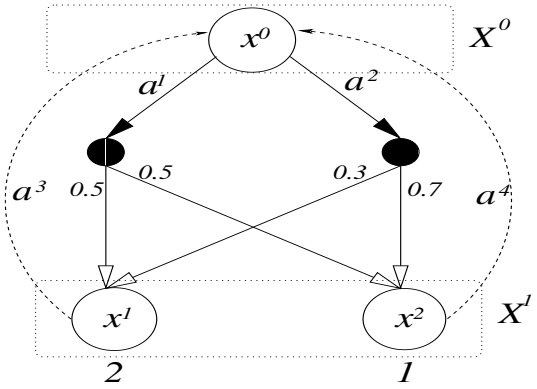


Figure 1: An example problem instance $\mathcal{E}$ where the state space X is partitioned into two layers, $X^0 = \{x^0\}$ and $X^1 = \{x^1, x^2\}$. The decisions associated with node X^0 are $\mathcal{A}(x^0) = \{\alpha^1, \alpha^2\}$, while every terminal node $x \in X^1$ has a single decision associated with it that leads back to state x^0. The transition probabilities for actions a^1 and a^2 are respectively $p(x^1; \alpha^1) = 0.5$, $p(x^2; \alpha^1) = 0.5$ and $p(x^1; \alpha^2) = 0.3$, $p(x^2; \alpha^2) = 0.7$; obviously, $p(x^0; \alpha^3) = p(x^0; \alpha^4) = 1$. Finally, the visitation requirement vector is defined by $\mathcal{N}_{x^1} = 2$, $\mathcal{N}_{x^2} = 1$.

x_0, in order to start another traversal. The entire operation terminates when all the node visitation requirements have been satisfied, i.e., $\mathcal{N}^c$ has been reduced to zero. An example problem instance is presented in Figure 1.

The induced stochastic shortest path problem Our intention is to determine an action selection scheme – or, a *policy* – π, that maps each tuple $(x, \mathcal{N}^c)$ to an action $\pi(x, \mathcal{N}^c) \in A(x)$ in a way that minimizes the expected number of graph traversals until $\mathcal{N}^c = \mathbf{0}$. This requirement can be further formalized through a Discrete Time MDP (DT-MDP), $\mathcal{M} = (S, A, t, c)$, where

- S is the finite set of *states*, identified with the tuples $(x, \mathcal{N}^c)$, where $x \in X$ and $\mathcal{N}^c \in \prod_{x \in X^L} \{0, \ldots, \mathcal{N}_x\}$.
- A is a set function defined on S that maps each state $s \in S$ to the finite, non-empty set $A(s)$, comprising all the *decisions / actions* that are feasible in s. More specifically, for $s = (x, \mathcal{N}^c)$, $A(s)$ coincides with $\mathcal{A}(x)$ as 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, i.e., a *partial* function defined on all triplets (s, a, s') with $a \in A(s)$, and with $t(s, a, s')$ being the probability to reach state s' from state s on decision a. More specifically, for $s = (x, \mathcal{N}^c)$, $a \in A(s)$,

$s' = (x', \mathcal{N}^{c'})$,

$$t(s,a,s') = \begin{cases} p(x';a), & \text{if } x \in X^l,\ l \in \{0,\ldots,L-1\},\ x' \in \bigcup_{k=l+1}^{L} X^k,\ \mathcal{N}^{c'} = \mathcal{N}^c; \\ 1, & \text{if } x \in X^L, x' = x^0,\ \mathcal{N}_x^{c'} = (\mathcal{N}_x^c - 1)^+,\ \mathcal{N}_y^{c'} = \mathcal{N}_y^c, \forall y \in X^L/\{x\}; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

- $c : S \longrightarrow \{0,1\}$ is the *cost function*, where for $s = (x, \mathcal{N}^c)$,

$$c(s) = \begin{cases} 1, & \text{if } x \in X^L \text{ and } \mathcal{N}^c \neq \mathbf{0}; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Notice that, under the considered cost function $c(\cdot)$, 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. For the purposes of the subsequent development, we shall represent this entire class of states with a single aggregate state, s^T, which we shall refer to as the problem *terminal state*; clearly, s^T is *absorbing* and *cost-free* under any policy π. Furthermore, the MDP state set S will be redefined to $S \equiv \{(x, \mathcal{N}^c) | \mathcal{N}^c \neq \mathbf{0}\} \cup \{s^T\}$, and the action, state transition and cost functions, A, t and c, will also be appropriately redefined to reflect the above aggregation. In particular, for the terminal state s^T, we define $A(s^T) = \{a^T\}$ with $t(s^T, a^T, s^T) = 1$; $t(s^T, a^T, s) = 0$, $\forall s \in S \backslash \{s^T\}$, and $c(s^T) = 0$. The redefinition of the remaining elements of A, t and c is straightforward and the relevant details are left to the reader.

In the above MDP modelling framework, a policy π that maps every state $s \in S$ to an action $\pi(s) \in A(s)$ is characterized as *stationary*. We are particularly interested in an *optimal* stationary 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. This optimality requirement for π^* can be formally characterized as follows: First, we define the expected total cost accumulated by the process when initialized at some state $s \in S$ and subsequently operated under some policy π, by

$$V^\pi(s) = E_\pi[\sum_{t=0}^{\infty} c(s_t) | s_0 = s] \tag{3}$$

where the expectation $E_\pi[\,]$ is taken over all possible process realizations under policy π. Next, we define

$$\pi_s^* = \arg\min_{\pi \in \Pi} V^\pi(s) \tag{4}$$

where Π is the set of all stationary policies.[1] Finally, we focus on $\pi^*_{s^0}$ where $s^0 \equiv (x^0, \mathcal{N})$, and we set

$$\pi^* \equiv \pi^*_{s^0} \quad \text{and} \quad V^* \equiv V^{\pi^*_{s^0}}(s^0) \tag{5}$$

The above specification of π^* and V^* brings the considered MDP problem to a particular class of MDP problems known as *stochastic shortest path (SSP)* problems [1]. For the resulting SSP problem to be well-defined, it remains to establish that (i) $V^* < \infty$ and (ii) the corresponding π^* is effectively computable. In order to derive these two results, we need to introduce the concept of a *proper* policy π:

Definition 1 *[1] For the considered SSP problem, a stationary policy π is said to be* proper *if and only if (iff) in the Markov chain induced by π, every state $s \in S\backslash\{s^T\}$ is connected to the terminal state s^T with an action path of positive probability. A stationary policy that is not proper will be said to be* improper*.*

The following proposition establishes the well-posed nature of the considered SSP problem and its proof can be found in the Appendix.

Proposition 1 *For the considered SSP problem, there exists at least one proper policy. Furthermore, for every improper policy π, there exists at least one state $s \in S\backslash\{s^T\}$ for which $V^\pi(s) = \infty$.*

The next theorem results immediately from the general SSP theory, in the light of Proposition 1; c.f. Proposition 2.1 in [1].

Theorem 1 *For the SSP formulation characterizing the problem considered in this work there exists a unique vector $V^*(s)$, $s \in S$, with $V^*(s^T) = 0$ and its remaining components, for $s \in S\backslash\{s^T\}$, satisfying the Bellman equation*

$$V^*(s) \;=\; \min_{a \in A(s)} \{c(s) + \sum_{s' \in S} t(s,a,s') \cdot V^*(s')\} \tag{6}$$

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

$$\pi^*(s) \;:=\; \arg\min_{a \in A(s)} \{c(s) + \sum_{s' \in S} t(s,a,s') \cdot V^*(s')\} \tag{7}$$

[1] It can be shown that for the considered problem, restriction to the class of stationary policies does not compromise the global optimality of the obtained solution; c.f. to Theorem 1 below.

The vector $V^*(s)$ is known as the *optimal value function* or the *optimal cost-to-go vector* for the considered SSP formulation. Each component of $V^*(s)$ expresses the expected total cost of initiating the underlying process at state $s \in S$ and subsequently following an optimal policy; in particular, $V^* = V^*(s^0)$. From a computational standpoint, $V^*(s)$ can be obtained through a number of approaches. Next, we focus on an approach based on linear programming that will also be useful in the subsequent developments presented in this document. We present the relevant result as a theorem, and we refer to [1] for the details of its derivation.

Theorem 2 *The optimal value vector $V^*(s),\ s \in S$, for the SSP formulation considered in this work is the optimal solution of the following linear program:*

$$\max \sum_{s \in S} V(s) \tag{8}$$

s.t.

$$\begin{aligned} & \forall s \in S \backslash \{s^T\},\ \forall a \in A(s), \\ V(s) \ &\leq\ c(s) + \sum_{s' \in S} t(s,a,s') \cdot V(s') \qquad (9) \\ V(s^T) \ &=\ 0 \qquad (10) \end{aligned}$$

Figure 2 exemplifies the structure of the SSP problem defined in this paragraph, by depicting the state transition diagram and the optimal policy for the SSP problem induced by the node visitation problem presented in Figure 1.

Complexity considerations and a spatially decomposing solution approach It should be clear from the definitions and the example provided in the previous paragraph, that the size of the state space S of the induced SSP problem is $|S| = |X| \cdot \prod_{x \in X^L} (\mathcal{N}_x + 1) - |X| + 1$. Hence, $|S|$ grows exponentially with respect to the size of $||\mathcal{N}||$, i.e., the number of the problem target nodes. This further implies that the computation – in fact, even the explicit enumeration – of the optimal value function $V^*(s),\ s \in S$, and the corresponding policy π^* will be a task of non-polynomial complexity with respect to the problem size $|\mathcal{E}|$. In particular, notice that the LP formulation of Theorem 2 will have $|S|$ variables and an even larger number of constraints. As a result, the LP-based solution approach delineated in Theorem 2 is severely limited by its computational complexity.

In the rest of this section we establish that the problem state space presents additional structure that enables the computation of the optimal value function $V^*(s),\ s \in S$, in an incremental

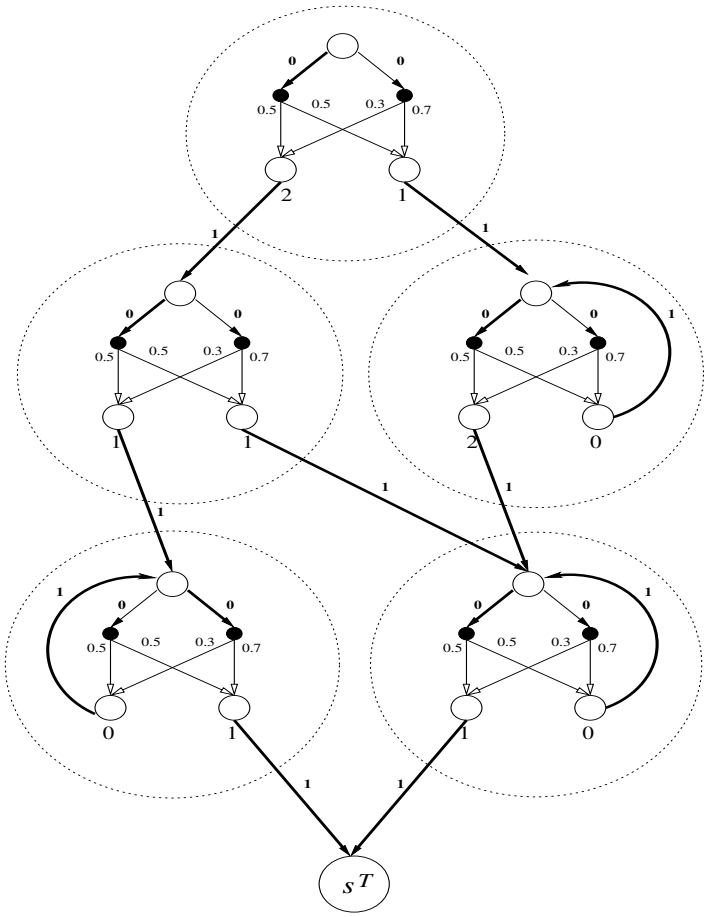


Figure 2: The State Transition Diagram for the stochastic shortest path problem induced by the problem instance $\mathcal{E}$ depicted in Figure 1. Each problem state corresponds to one of the non-filled (white) nodes in the depicted graph, and it is defined by (i) the position of this node in the underlying acyclic graph that defines the problem instance, and (ii) the vector of the remaining visitation requirements. The values annotated next to the arcs representing the available actions at every state indicate the corresponding immediate cost. By associating a non-zero, unit cost only with the actions / transitions emanating from states where the first component is a leaf node, it is ensured that the cost accumulated over any sample path leading from the initial state s^0 to the terminal state s^T, is equal to the corresponding number of graph traversals experienced by the control agent. Finally, the reader can verify that the optimal policy for the considered problem instance is defined by the emboldened arcs in this graph, and that $V^* = 4.357$.

fashion, by solving a sequence of LP formulations, each containing a number of variables and constraints that are polynomially related to $|\mathcal{E}|$. More specifically, the decomposing solution approach presented in this paragraph is based on the following key observations:[2] First, notice that, by the definition of the DT-MDP $\mathcal{M}$, every cycle appearing in the state space S involves states $s \in S$ that have the same vector $\mathcal{N}^c$ as their second component. Furthermore, every transition from a state $s = (x, \mathcal{N}^c)$ with $|\mathcal{N}^c| > 0$ leads to another state $s' = (x', \mathcal{N}^c)$ or to a state $s'' = (x^0, \mathcal{N}^{c''})$ with $|\mathcal{N}^{c''}| = |\mathcal{N}^c| - 1$. When combined with the structure of the Bellman equation, that characterizes the optimal value function $V^*(s)$ (c.f. Equation 6), the above two observations imply that, for every state $s = (x, \mathcal{N}^c)$ with $|\mathcal{N}^c| \geq 1$, (i) $V^*(s)$ is completely defined by the optimal values $V^*(s')$ for $s' \in \{(x, \mathcal{N}^c) | x \in X\} \cup \{(x^0, \mathcal{N}^{c'}) |\ |\mathcal{N}^{c'}| = |\mathcal{N}^c| - 1\}$, (ii) but itself has no impact on the determination of the optimal values $V^*(s')$ for states $s' \in \{(x^0, \mathcal{N}^{c'}) |\ |\mathcal{N}^{c'}| = |\mathcal{N}^c| - 1\}$. Hence, the LP of Theorem 2 can be solved incrementally through an iterative procedure that computes the value function for the subsets of states corresponding to distinct vectors $\mathcal{N}^c$ one at a time, starting with the state subsets with $|\mathcal{N}^c| = 1$ and proceeding to the state subset corresponding to $\mathcal{N}^c = \mathcal{N}$; a formal characterization of this procedure is provided in Figure 3.

Clearly, each of the LP's solved under the above solution approach is polynomially sized with respect to $|\mathcal{E}|$. However, this approach is still limited by the fact that the total number of linear programs to be solved is equal to $\prod_{x \in ||\mathcal{N}||} (\mathcal{N}_x + 1)$, which remains a non-polynomial quantity with respect to $|\mathcal{E}|$. For this reason, in Section 3 we also propose an alternative solution approach that in general will lead to a sub-optimal policy, but, both, the derivation and implementation of this policy will be of polynomial complexity with respect to the problem size $|\mathcal{E}|$. Furthermore, we shall show that the value of this policy converges to the value of the optimal policy as the target visitation requirements grow uniformly to infinity.

3 A computationally efficient and asymptotically optimal policy

The main contribution of this section is a randomized policy for the MDP problem defined in Section 2, that is of polynomial complexity with respect to the problem size $|\mathcal{E}|$, and the ratio of its value to the value V^*, of the optimal policy π^*, converges to unity, as the non-zero node visitation requirements grow uniformly to infinity. The definition and the properties of this policy rely heavily on the LP formulation of a surrogate deterministic optimization problem

[2] The reader is referred to Figure 2 for a more concrete demonstration of these observations.

Computing the Optimal Value Function through Spatial Decomposition

$V^*(s^T) := 0$

for $i := 1$ to $|\mathcal{N}|$

  if $(i < |\mathcal{N}|)$

    $\Omega := \{\mathcal{N}^c |\ |\mathcal{N}^c| = i\}$

  else

    $\Omega := \{\mathcal{N}\}$

  for $\mathcal{N}^c \in \Omega$

    Solve the LP obtained from the LP of Theorem 2 when the variable

    and the constraint sets are restricted to those corresponding to $s \in \{(x, \mathcal{N}^c) | x \in X\}$,

    and all the remaining variables $V(s)$ in the aforementioned constraints are

    substituted by the values $V^*(s)$ obtained from the solution of the earlier subproblems.

Return as $V^*(s),\ s \in S$, the vector obtained from the concatenation of the solutions of the aforementioned LP's.

Figure 3: An iterative algorithm for computing the optimal value function of the considered shortest path problem, through spatial decomposition

that in the following will be referred to as the *"relaxing LP"*. Hence, the first part of this section introduces the relaxing LP formulation and the underlying optimization problem, and it establishes that this formulation provides a lower bound for V^*. Subsequently, the second part employs the optimal solution of the relaxing LP in order to define the aforementioned randomized policy, and proves its asymptotic optimality.

3.1 The "Relaxing LP" and its relationship to the optimal value of the SSP formulation

The relaxing LP is the analytical characterization of the following problem: Consider the acyclic graph $\mathcal{G}$ introduced in Section 2, and assume that a certain amount of fluid is pumped from the root node, x^0, of $\mathcal{G}$ to its terminal nodes, $x \in X^L$. At each non-terminal node, $x \in X^l,\ l = 0, 1 \ldots, L-1$, the incoming flow is conveyed to the emanating arcs corresponding to the various actions $a \in \mathcal{A}(x)$ according to a routing scheme to be determined by the considered formulation. On the other hand, the flow directed to an arc $a \in \mathcal{A}(x),\ x \in \bigcup_{l=0}^{L-1} X^l$, is distributed to the nodes $x' \in \mathcal{S}(a)$ according to the proportions defined by the probability function $p(x'; a), x' \in \mathcal{S}(a)$. We want to determine the fluid volume to be routed through each

arc $a \in \mathcal{A}(x),\ x \in \bigcup_{l=0}^{L-1} X^l$, so that each terminal node $x \in X^L$ receives a fluid volume at least equal to $\mathcal{N}_x$, while the total amount of fluid induced into graph $\mathcal{G}$ through its root node x^0 is minimized. Letting χ_a denote the fluid volume routed through arc $a \in \mathcal{A}(x),\ x \in \bigcup_{l=0}^{L-1} X^l$, the above problem can be expressed by the following LP formulation:

$$\min \sum_{a \in \mathcal{A}(x^0)} \chi_a \tag{11}$$

s.t.

$$\begin{aligned} & & & \forall x \in X \backslash (\{x^0\} \cup X^L), \\ \sum_{a: x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a & = & \sum_{a \in \mathcal{A}(x)} \chi_a & \qquad (12) \\ & & & \forall x \in X^L, \\ \sum_{a: x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a & \geq & \mathcal{N}_x & \qquad (13) \\ & & & \forall a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x), \\ \chi_a & \geq & 0 & \qquad (14) \end{aligned}$$

As it was already pointed out, we shall refer to the LP formulation of Equations 11–14 as the *"relaxing LP"*, and we shall denote the optimal value of this formulation by V^*_{rel}. In the sequel we shall establish that V^*_{rel} is a lower bound for V^*. However, the establishment of this result will employ an analytical characterization of V^* that is based on a variant of the LP formulation introduced in Theorem 2. More specifically, this new formulation computes $V^* = V^*(s^0)$ through the following LP

$$\max\ V(s^0) \tag{15}$$

s.t.

$$\begin{aligned} & & \forall s \in S \backslash \{s^T\},\ \forall a \in A(s), \\ V(s) & \leq c(s) + \sum_{s' \in S \backslash \{s^T\}} t(s,a,s') \cdot V(s') & \qquad (16) \end{aligned}$$

where (i) the zero-priced variable $V(s^T)$ has been eliminated and (ii) the objective function of Equation 8 has been substituted with the objective function of Equation 15. The performed substitution is legitimate because it is well-known in the relevant MDP theory that the SSP optimal value function $V^*(s),\ s \in S$, is the *componentwise maximal* vector that satisfies the constraint of Equation 9 for $V(s^T) = 0$. Furthermore, instead of computing $V^*(s^0)$ directly

through the LP of Equations 15–16, in the subsequent discussion will shall focus on the *dual* LP of this formulation [4]. Letting $q(s,a)$ denote the dual variable corresponding to the primal constraint for the state-action pair $(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$, and also using the notation $x(s)$ in order to denote the first component of any state $s = (x, \mathcal{N}^c)$, this dual LP can be written as follows (c.f. [4]):

$$\min \sum_{s \in S\backslash\{s^T\}: x(s) \in X^L} \sum_{a \in A(s)} q(s,a) \tag{17}$$

s.t.

$$\begin{aligned} & \forall s \in S\backslash\{s^T\}, & (18) \\ \sum_{a \in A(s)} q(s,a) &= \mathbf{1}_{\{s=s^0\}} + \sum_{s' \in S\backslash\{s^T\}} \sum_{a \in A(s')} t(s',a,s) \cdot q(s',a) \\ & \forall s \in S\backslash\{s^T\},\ \forall a \in A(s), \\ q(s,a) &\geq 0 & (19) \end{aligned}$$

An optimal solution of the LP formulation of Equations 17–19 will be denoted by $q^*(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$. It is well-known from LP duality theory [4] that

$$\sum_{s \in S\backslash\{s^T\}: x(s) \in X^L} \sum_{a \in A(s)} q^*(s,a) = V^*(s^0) \tag{20}$$

In addition, any feasible solution $q(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$, of the dual LP formulation admits a flow interpretation in the state transition diagram (STD) defined by the MDP state set S and the corresponding action sets $A(s),\ s \in S$. More specifically, under this interpretation, any feasible solution $q(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$, of the dual LP formulation defines a flow pattern that transfers a unit flow entering the aforementioned STD at the initial state s^0 to the terminal state s^T. In this context, the constraint of Equation 18 expresses a flow balance requirement, while the objective function of Equation 17 measures the flow that is routed through the arcs corresponding to actions $a \in A(s)$ with $s \in S\backslash\{s^T\}$ and $x(s) \in X^L$. Next we shall employ this flow interpretation of the feasible solutions of the dual LP formulation of the Equations 17–19 in order to prove the following theorem:

Theorem 3 *Under the above definitions,* $V^*_{rel} \leq V^*$.

Proof Equation 20 implies that in order to prove the result of Theorem 3, it suffices to show that (i) every feasible solution $q(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$, for the LP formulation of Equations 17–19, induces a feasible solution $\chi_a,\ a \in \bigcup_{x \in X\backslash X^L} \mathcal{A}(x)$, for the relaxing LP,

and (ii) the corresponding objective values are equal. Hence, consider such a feasible solution $q(s,a),\ s \in S\backslash\{s^T\},\ a \in A(s)$, for the dual LP formulation of Equations 17–19, and define

$$\chi_a \equiv \sum_{s \in S\backslash\{s^T\}: a \in A(s)} q(s,a), \quad \forall a \in \bigcup_{x \in X\backslash X^L} \mathcal{A}(x) \tag{21}$$

In the remaining part of this proof we shall show that the vector $\{\chi_a\}$ defined by Equation 21 satisfies the aforestated requirements, when considered as a solution to the relaxing LP.

Clearly, Constraint 14 is immediately satisfied by Constraint 19 and the definition of $\{\chi_a\}$. Next we prove the feasibility of $\{\chi_a\}$ with respect to Constraint 12. Hence, consider a node $x \in X\backslash(\{x^0\} \cup X^L)$. Then it holds that:

$$\begin{aligned}
\sum_{a \in \mathcal{A}(x)} \chi_a &= \sum_{a \in \mathcal{A}(x)} \sum_{s \in S\backslash\{s^T\}: a \in A(s)} q(s,a) \\
&= \sum_{s \in S\backslash\{s^T\}: x(s) = x} \sum_{a \in A(s)} q(s,a) \\
&= \sum_{s \in S\backslash\{s^T\}: x(s) = x} \sum_{s' \in S\backslash\{s^T\}} \sum_{a \in A(s')} t(s',a,s) \cdot q(s',a) \\
&= \sum_{a: x \in \mathcal{S}(a)} p(x;a) \cdot \sum_{s' \in S\backslash\{s^T\}: a \in A(s')} q(s',a) \\
&= \sum_{a: x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a
\end{aligned}$$

The first equality above results from Eq. 21, the second from term rearrangement, the third from Eq. 18, the fourth from the definition of the function t and term rearrangement, and the last from Eq. 21.

To prove the satisfaction of Constraint 13 by the vector $\{\chi_a\}$, first notice that this constraint is trivially satisfied for all non-target nodes $x \in X^L$. Hence, consider a node $x \in X^L$ with $\mathcal{N}_x > 0$. Then, by working as in the proof of the validity of Constraint 12, we can easily establish that

$$\sum_{a: x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a = \sum_{s \in S\backslash\{s^T\}: x(s) = x} \sum_{a \in A(s)} q(s,a) \tag{22}$$

Next consider the arc set $\mathcal{C}_x(\mathcal{N}_x)$, consisting of all the arcs in the STD defined by the state set S and the action sets $A(s),\ s \in S$, that lead from any state $s \in S_x(\mathcal{N}_x) \equiv \{(x, \mathcal{N}^c) : \mathcal{N}^c_x = \mathcal{N}_x\}$ to the resultant state $s' = (x^0, \mathcal{N}^c - \mathbf{1}_x)$, where $\mathbf{1}_x$ denotes the unit vector of dimensionality $|X^L|$ and with the non-zero component corresponding to node x.[3] Clearly, since x is a target node, $\mathcal{C}_x(\mathcal{N}_x)$ is non-empty. Furthermore, since this set aggregates all the possible transitions

[3]The reader is referred to Figure 4 for a more concrete visualization of the concepts and arguments related to this part of the proof.

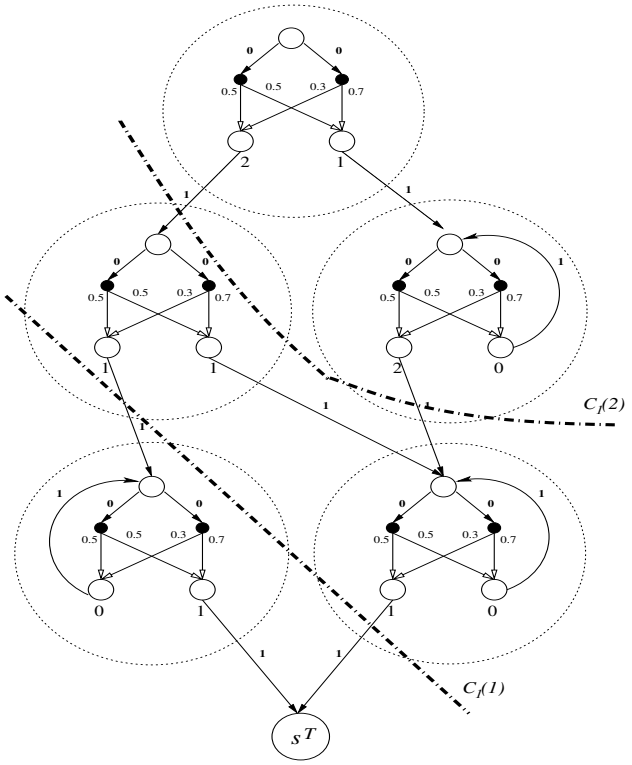


Figure 4: The STD *"cuts"* $\mathcal{C}_1(1)$ and $\mathcal{C}_1(2)$ defined by the target leaf node x^1 in the optimal node visitation problem of Figure 1

through which the visitation requirements for x are reduced from $\mathcal{N}_x$ to $\mathcal{N}_x - 1$, it defines a *cut* on the underlying graph defined by S and $A(s)$, $s \in S$. This last observation when combined with the fact that $\{q(s,a)\}$ defines a flow that conveys a unit load from state s^0 to state s^T, imply that

$$\sum_{(s,a)\in\mathcal{C}_x(\mathcal{N}_x)} q(s,a) = 1 \tag{23}$$

In the same way, we can define the arc sets $\mathcal{C}_x(\mathcal{N}_x - k)$, $k \in \{1,\ldots,\mathcal{N}_x - 1\}$, each consisting of all the arcs that lead from any state $s \in S_x(\mathcal{N}_x - k) \equiv \{(x,\mathcal{N}^c) : \mathcal{N}_x^c = \mathcal{N}_x - k\}$ to the state $s' = (x^0, \mathcal{N}^c - \mathbf{1}_x)$, and establish that

$$\sum_{(s,a)\in\mathcal{C}_x(\mathcal{N}_x-k)} q(s,a) = 1, \quad \forall k \in \{1,\ldots,\mathcal{N}_x - 1\} \tag{24}$$

But then, the satisfaction of Constraint 13 results immediately from the fact that each of the summations appearing in Equations 23 and 24 is subsumed in the double summation that appears in the right-hand-side of Equation 22.

It remains to show that

$$\sum_{a\in\mathcal{A}(x^0)} \chi_a = \sum_{s\in S\setminus\{s^T\}:x(s)\in X^L} \sum_{a\in A(s)} q(s,a)$$

The validity of this equation is established as follows:

$$\begin{aligned}
\sum_{a\in\mathcal{A}(x^0)} \chi_a &= \sum_{s\in S\setminus\{s^T\}:x(s)=x^0} \sum_{a\in A(s)} q(s,a) \\
&= \sum_{s\in S\setminus\{s^T,s^0\}:x(s)=x^0} \sum_{a\in A(s)} q(s,a) + \sum_{a\in A(s^0)} q(s^0,a) \\
&= \sum_{s\in S\setminus\{s^T,s^0\}:x(s)=x^0} \sum_{s'\in S\setminus\{s^T\}} \sum_{a\in A(s')} t(s',a,s)\cdot q(s',a) + 1 + \\
&\quad \sum_{s'\in S\setminus\{s^T\}} \sum_{a\in A(s')} t(s',a,s^0)\cdot q(s',a) \\
&= 1 + \sum_{s\in S\setminus\{s^T\}:x(s)=x^0} \sum_{s'\in S\setminus\{s^T\}} \sum_{a\in A(s')} t(s',a,s)\cdot q(s',a) \\
&= 1 + \sum_{s\in S:x(s)=x^0} \sum_{s'\in S\setminus\{s^T\}} \sum_{a\in A(s')} t(s',a,s)\cdot q(s',a) - 1 \\
&= \sum_{s\in S\setminus\{s^T\}:x(s)\in X^L} \sum_{a\in A(s)} q(s,a)
\end{aligned}$$

The first equality above can be derived as in the proof of Constraint 12, the third equality results from Eq. 18, the fifth equality results from the fact that $\sum_{s'\in S\setminus\{s^T\}} \sum_{a\in A(s')} t(s',a,(x^0,\mathbf{0}))\cdot q(s',a) = 1$, and the last from the definition of function t. ∎

3.2 The proposed randomized policy and its asymptotic optimality

In this section we introduce a randomized policy for the MDP problem defined in Section 2 and establish its asymptotic optimality.[4] The definition of this policy relies on the optimal solution of the relaxing LP, introduced in Section 3.1. In particular, given an optimal solution $\{\chi_a^* | a \in \bigcup_{x\in X\setminus X^L} \mathcal{A}(x)\}$ of this LP, we determine a policy π that assigns to a state $s=(x,\mathcal{N}^c)$ with $x \in X\setminus X^L$ and $\sum_{a\in\mathcal{A}(x)} \chi_a^* > 0$, an action $\pi(x,\mathcal{N}^c) \in A(s)$ according to the probability distribution

$$P(\pi(x,\mathcal{N}^c)=a) = \frac{\chi_a^*}{\sum_{a\in\mathcal{A}(x)} \chi_a^*}, \quad a\in\mathcal{A}(x). \tag{25}$$

[4]We remind the reader that a randomized policy for the considered MDP problem is an action selection scheme where, at each state $s=(x,\mathcal{N}^c)$, an action $a\in A(s)$ is selected according to a probability distribution supported on the set $A(s)$.

On the other hand, for states $s = (x, \mathcal{N}^c)$ with $x \in X \backslash X^L$ and $\sum_{a \in \mathcal{A}(x)} \chi_a^* = 0$, the policy is indeterminate. Finally, for states $s = (x, \mathcal{N}^c)$, $x \in X^L$, the policy executes the unique transition $a \in A(s)$ with probability one.

Clearly, the deployment and execution of the aforestated policy π is of polynomial complexity with respect to the problem size $|\mathcal{E}|$, since this complexity is determined by (i) the solution of the relaxing LP, (ii) the computation and storage in a pertinent data structure of the action selection distributions induced by the optimal solution $\{\chi_a^* | a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x)\}$, for all nodes $x \in X \backslash X^L$ with $\sum_{a \in \mathcal{A}(x)} \chi_a^* > 0$, and (iii) the reference of these distributions every time that the underlying process enters a state $s = (x, \mathcal{N}^c)$ with $x \in X \backslash X^L$ and $\sum_{a \in \mathcal{A}(x)} \chi_a^* > 0$. The next proposition also establishes a notion of properness for the aforestated policy, and its proof can be found in the Appendix.

Proposition 2 *The proposed randomized policy, π, has the following two properties:*

i. Every state $s = (x, \mathcal{N}^c)$ with $x \in X \backslash X^L$ and $\sum_{a \in \mathcal{A}(x)} \chi_a^* = 0$ is unreachable under policy π.

ii. For every state $s \in S \backslash \{s^T\}$ that is reachable under policy π, there exists an action path leading from s to the terminal state s^T with positive probability.

In the remaining part of this section we establish the asymptotic optimality of the considered randomized policy π; in particular, we show that the ratio of the policy value, $V^\pi \equiv V^\pi(s^0)$, to the value V^*, of the optimal policy π^*, converges to unity, as the non-zero node visitation requirements grow uniformly to infinity. In order to formally state and prove this convergence, for any given problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$, we shall consider the entire problem sequence, $\{\mathcal{E}(n)\}$, obtained by replacing the visitation requirement vector, $\mathcal{N}$, with $n \cdot \mathcal{N}$, $n \in Z^+$, and letting $n \to \infty$. Also, we shall let (i) $\{V^*(n)\}$ denote the sequence of the optimal expected total costs for the corresponding problem instances $\mathcal{E}(n)$, (ii) $\{\chi_a^*(n) | a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x)\}$ and $\{V_{rel}^*(n)\}$ denote respectively the sequences of the optimal solutions and the optimal objective values for the corresponding relaxing LP's, (iii) $\{\pi(n)\}$ denote the sequence of the randomized policies defined for the various elements of $\{\mathcal{E}(n)\}$ by the corresponding elements of $\{\chi_a^*(n)\}$, and (iv) $\{\widehat{V}^\pi(n)\}$ denote the sequence of the random costs incurred by each randomized policy $\pi(n)$ when exercised upon its corresponding problem instance $\mathcal{E}(n)$. We already know from Theorem 3 that $V^\pi(n) \equiv E_\pi[\widehat{V}^\pi(n)] \geq V^*(n) \geq V_{rel}^*(n)$, $\forall n > 0$. The following series of lemmata establishes that the ratio $\widehat{V}^\pi(n)/V_{rel}^*(n)$ converges almost surely to unity, as $n \to \infty$.

Lemma 1 *Consider the problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$ and an optimal solution, $\{\chi_a^* | a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x)\}$, of the corresponding relaxing LP, defined by Equations 11–14. Then, for every*

node $x \in X$, the process defined by the initial state $s' = (x^0, \mathcal{N}^c)$ and the randomized policy π, that is induced by $\{\chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ as described at the beginning of this section, will reach the state $(x, \mathcal{N}^c)$ before revisiting the set $\{s \in S | x(s) = x^0\}$ with probability

$$\mathcal{P}_x = \frac{\sum_{a:x \in \mathcal{S}(a)} p(x;a) \cdot \chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*}. \tag{26}$$

The proof of this lemma is through a straightforward induction on the layer index, l, and it can be found in the Appendix.

Lemma 2 *Consider the problem sequence, $\{\mathcal{E}(n)\}$, that is induced by a problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$, through the scaling of the visitation requirement vector, $\mathcal{N}$, by a factor $n \in Z^+$. Then, for all $n \in Z^+$,*

$$V_{rel}^*(n) = n \cdot \max_{x:\mathcal{N}_x > 0} \{\frac{\mathcal{N}_x}{\mathcal{P}_x}\} \tag{27}$$

Proof First we prove the validity of Equation 27 for $n = 1$. Since $\{\chi_a^*(1) | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\} \equiv \{\chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ is an optimal solution of the relaxing LP, it must hold: $\sum_{a:x \in S(a)} \chi_a^*(1) \cdot p(x;a) \geq \mathcal{N}_x, \forall x \in X^L$. When combined with Equation 26, this last inequality implies that $\mathcal{P}_x \cdot (\sum_{a \in \mathcal{A}(x^0)} \chi_a^*(1)) \geq \mathcal{N}_x, \forall x \in X^L$. Taking into account that (i) since $\{\chi_a^*\}$ is an optimal solution of the relaxing LP, at least one of the Constraints 13 must hold as equality, (ii) $V_{rel}^* = \sum_{a \in \mathcal{A}(x^0)} \chi_a^*$, and (iii) $\mathcal{P}_x > 0, \ \forall x \in X^L$ with $\mathcal{N}_x > 0$, we finally get

$$V_{rel}^*(1) \equiv V_{rel}^* = \max_{x:\mathcal{N}_x > 0} \{\frac{\mathcal{N}_x}{\mathcal{P}_x}\} \tag{28}$$

Next fix an $n > 1$. We shall refer to the relaxing LP corresponding to the problem instance $\mathcal{E}(n)$ as LP(n), and to the relaxing LP of the original problem instance $\mathcal{E}$ as LP(1). Notice that the dual of LP(n) has the same feasible region as the dual of LP(1), while the objective function of the former is equal to the objective function of the latter multiplied by n. Therefore, the optimal objective value of the dual of LP(n) is equal to the objective value of the dual of LP(1) multiplied by n. This last result, when combined with LP duality theory [4] and Equation 28, imply that $V_{rel}^*(n) = n \cdot \max_{x:\mathcal{N}_x > 0} \{\frac{\mathcal{N}_x}{\mathcal{P}_x}\}$, and establish the validity of Equation 27 for every $n > 0$. ∎

Lemma 3 *Let $\{\chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ be an optimal solution to the relaxing LP corresponding to a problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$. Then, $\{\chi_a^*(n) \equiv n \cdot \chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ is an*

optimal solution to the relaxing LP corresponding to the problem instance $\mathcal{E}(n)$, for all $n \in Z^+$.

Proof The definition of the problem instance $\mathcal{E}(n)$ implies that the vector $\{\chi_a^*(n) \equiv n \cdot \chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ is a feasible solution for the corresponding relaxing LP. Furthermore, the objective value obtained by plugging the considered vector, $\{\chi_a^*(n)\}$, to the expression of Equation 11 is equal to $n \cdot V_{rel}^*$, which, by the result of Lemma 2, is equal to $V_{rel}^*(n)$. ∎

An immediate implication of Lemma 3 is that the randomized policies $\pi(n)$, for $n > 1$, are identical to the policy $\pi \equiv \pi(1)$, that is induced by the optimal solution $\{\chi_a^* | a \in \bigcup_{x \in X \setminus X^L} \mathcal{A}(x)\}$ of the relaxing LP for the original problem instance, $\mathcal{E}$. The next lemma employs this result, together with the results of Lemmas 1–3, in order to state and prove the key result of this section.

Lemma 4 *Consider a problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$ and the sequence of problem instances, $\{\mathcal{E}(n)\}$, that is induced by $\mathcal{E}$ through the scaling of the visitation requirement vector, $\mathcal{N}$, by $n \in Z^+$. Also, let (i) $\{V_{rel}^*(n)\}$ denote the sequence of the optimal values for the corresponding relaxing LP's; (ii) π denote the randomized policy defined by an optimal solution of the relaxing LP for $\mathcal{E}$; and (iii) $\{\widehat{V}^\pi(n)\}$ denote the sequence of the random costs incurred by the application of the randomized policy π to the problem instances $\mathcal{E}(n)$. Then, for $n \to \infty$,*

$$\frac{\widehat{V}^\pi(n)}{V_{rel}^*(n)} \xrightarrow{a.s.} 1 \tag{29}$$

Proof Consider the application of the randomized policy π on some problem instance $\mathcal{E}(n)$, and, for each terminal node $x \in X^L$ and each $i = 1, \ldots, \widehat{V}^\pi(n)$, define the random variable $I_i^x = 1$, if the process visits node x during its i-th traversal of the graph $\mathcal{G}$; 0, otherwise. Then, the number of times that the process visits a terminal node $x \in X^L$ before its termination, can be expressed as $\sum_{i=1}^{\widehat{V}^\pi(n)} I_i^x$. By the problem definition, $\sum_{i=1}^{\widehat{V}^\pi(n)} I_i^x \geq n \cdot \mathcal{N}_x, \forall x \in X^L$, or equivalently, $\widehat{V}^\pi(n) \cdot \frac{\sum_{i=1}^{\widehat{V}^\pi(n)} I_i^x}{\widehat{V}^\pi(n)} \geq n \cdot \mathcal{N}_x, \forall x \in X^L$. Since the constraint attained last must be holding as equality, we obtain

$$\frac{\widehat{V}^\pi(n)}{n} = \max_{x: \mathcal{N}_x > 0} \{ \frac{\mathcal{N}_x}{\frac{\sum_{i=1}^{\widehat{V}^\pi(n)} I_i^x}{\widehat{V}^\pi(n)}} \} \tag{30}$$

Observe that $\widehat{V}^\pi(n) \geq n \cdot \sum_{x \in X^L} \mathcal{N}_x$ a.s., which implies that $\widehat{V}^\pi(n) \xrightarrow{a.s.} \infty$ as $n \to \infty$. But then, the result of Lemma 1, when combined with the definition of I_i^x and the Strong Law of

Large Numbers [2], imply that

$$\frac{\sum_{i=1}^{\widehat{V}^{\pi}(n)} I_i^x}{\widehat{V}^{\pi}(n)} \xrightarrow{a.s.} \mathcal{P}_x \tag{31}$$

as $n \to \infty$. Subsequently, Equations 30, 31, and an application of the Continuous Mapping Theorem [2] imply that

$$\frac{\widehat{V}^{\pi}(n)}{n} \xrightarrow{a.s.} \max_{x:\mathcal{N}_x>0} \{\frac{\mathcal{N}_x}{\mathcal{P}_x}\} \tag{32}$$

as $n \to \infty$. Equation 32, when combined with Lemma 2, imply that

$$\frac{\widehat{V}^{\pi}(n)}{V^*_{rel}(n)} = \frac{\widehat{V}^{\pi}(n)}{n \cdot V^*_{rel}} \xrightarrow{a.s.} 1 \tag{33}$$

as $n \to \infty$, and conclude the proof. ∎

Finally, the next theorem builds upon Lemma 4 and some technical results regarding the uniform integrability of random variables, in order to formally establish the asymptotic optimality of the proposed randomized policy π, as the node visitation requirements grow uniformly to infinity.

Theorem 4 *Consider a problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$ and the sequence of problem instances, $\{\mathcal{E}(n)\}$, that is induced by $\mathcal{E}$ through the scaling of the visitation requirement vector, $\mathcal{N}$, by $n \in Z^+$. Also, let (i) $\{V^*(n)\}$ denote the sequence of the corresponding optimal expected costs; (ii) π denote the randomized policy defined by an optimal solution of the relaxing LP for $\mathcal{E}$; and (iii) $\{\widehat{V}^{\pi}(n)\}$ denote the sequence of the random costs incurred by the application of the randomized policy π to the problem instances $\mathcal{E}(n)$. Then, for $n \to \infty$,*

$$\frac{\widehat{V}^{\pi}(n)}{V^*(n)} \xrightarrow{a.s.} 1 \tag{34}$$

Proof Since

$$\forall n \in Z^+, \quad V^*_{rel}(n) \leq V^*(n) \leq V^{\pi}(n)$$

and, from Lemma 4,

$$\frac{V^*_{rel}(n)}{\widehat{V}^{\pi}(n)} \xrightarrow{a.s.} 1$$

as $n \to \infty$, it suffices to show that

$$\frac{V^{\pi}(n)}{\widehat{V}^{\pi}(n)} = \frac{V^{\pi}(n)/n}{\widehat{V}^{\pi}(n)/n} \xrightarrow{a.s.} 1 \tag{35}$$

as $n \to \infty$. Furthermore, since $V^{\pi}(n) = E_{\pi}[\widehat{V}^{\pi}(n)]$, Eq. 32 implies that a sufficient condition for Eq. 35 to hold, is that

$$\lim_{n\to\infty} E_{\pi}[\widehat{V}^{\pi}(n)/n] = E_{\pi}[\lim_{n\to\infty} \widehat{V}^{\pi}(n)/n] \tag{36}$$

This last result can be established as follows: Let $x^l = \arg\max_{x\in X^L:\mathcal{N}_x>0}\{\frac{\mathcal{N}_x}{\mathcal{P}_x}\}$, i.e., x^l denotes the *"most difficult"* target node under the randomized policy π. Also, for any given $n \in Z^+$, let $\Psi(n)$ be the random cost resulting from the application of the policy π on the problem instance $\mathcal{E}^l(n)$, that is obtained from $\mathcal{E}(n)$ by setting $\mathcal{N}_x = 0$ for $x \neq x^l$. Clearly,

$$\widehat{V}^{\pi}(n) \leq_{st} |X^L| \cdot \Psi(n) \tag{37}$$

since $|X^L| \cdot \Psi(n)$ is a stochastic upper bound for the performance that is attained by a policy that seeks to satisfy the posed visitation requirements one leaf node at a time, while routing tokens according to the routing probabilities employed by policy π.

By definition, $\Psi(n) - 1$ follows a negative binomial distribution with mean

$$E[\Psi(n) - 1] = E[\Psi(n)] - 1 = n\mathcal{N}_{x^l}\frac{1-\mathcal{P}_{x^l}}{\mathcal{P}_{x^l}} \tag{38}$$

and variance

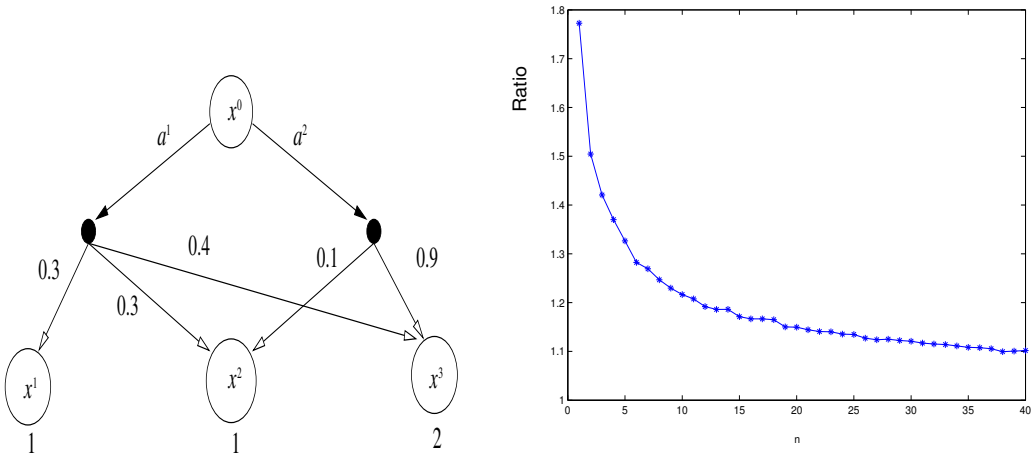
$$Var[\Psi(n) - 1] = Var[\Psi(n)] = n\mathcal{N}_{x^l}\frac{1-\mathcal{P}_{x^l}}{\mathcal{P}^2_{x^l}} \tag{39}$$

Therefore,

$$\begin{aligned}
\sup_n\left\{E[\frac{\Psi^2(n)}{n^2}]\right\} &= \sup_n\left\{\frac{1}{n^2}(Var[\Psi(n)] + E[\Psi(n)]^2)\right\} \\
&= \sup_n\left\{\frac{1}{n}\mathcal{N}_{x^l}\frac{1-\mathcal{P}_{x^l}}{\mathcal{P}^2_{x^l}} + \frac{1}{n^2}(n\mathcal{N}_{x^l}\frac{1-\mathcal{P}_{x^l}}{\mathcal{P}_{x^l}} + 1)^2\right\} \\
&< \infty
\end{aligned} \tag{40}$$

But then, $\Psi(n)/n$ are uniformly integrable (c.f., [3], pg. 338), which combined with Equation 37, implies the uniform integrability of the random variables $\widehat{V}^{\pi}(n)/n$. Finally, the desired result of Equation 36 is obtained from the uniform integrability of $\widehat{V}^{\pi}(n)/n$ and Equation 32, when combined with Theorem 25.12 of [3] (c.f. pg. 338). ∎

As an example of the result of Theorem 4, consider the problem instance $\mathcal{E}$ depicted in Figure 5(a), where the node set X is partitioned into two layers $X^0 = x^0$, $X^1 = X^L = \{x^1, x^2, x^3\}$, and the decisions associated with the non-terminal node x^0 are $\mathcal{A}(x^0) = \{\alpha^1, \alpha^2\}$. The transition function is given by $p(x^1;\alpha^1) = 0.3$, $p(x^2;\alpha^1) = 0.3$, $p(x^3;\alpha^1) = 0.4$, $p(x^2;\alpha^2) = 0.1$, $p(x^3;\alpha^2) = 0.9$, and $p(x^0, \alpha(x^l)) = 1,\ \forall x^l \in X^1$. Finally the visitation requirement vector



(a) **(b)**

Figure 5: A demonstration of the result of Theorem 4

is defined by $\mathcal{N}_{x^1} = 1$, $\mathcal{N}_{x^2} = 1$, $\mathcal{N}_{x^3} = 2$. The graph illustrated in Figure 5 (b) demonstrates the ratio of the random performance of the randomized policy π, $\widehat{V}^{\pi}(n)$, and the lower bound $V^*_{rel}(n)$ returned by the relaxing LP, as the visitation requirement vector is scaled uniformly by a factor n. It is obvious from the plot that as the value of n increases, this ratio tends to one.

4 Establishing an upper bound for V^*

In order to develop an upper bound, $\hat{V}$, for the optimal value, V^*, of any given problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$, it suffices to compute the expected total cost, $V^{\hat{\pi}}(s^0)$, for any viable policy $\hat{\pi}$. In the sequel, we shall focus on the particular policy $\hat{\pi}$ that seeks to satisfy the posed visitation requirements for each terminal node sequentially, one node at a time. More specifically, the policy $\hat{\pi}$ first orders the nodes $x \in X^L$ with $\mathcal{N}_x > 0$ in some arbitrary sequence, and subsequently, it iteratively selects the next node in the sequence and tries to satisfy its visitation requirements while ignoring all the visits to any other terminal node. Furthermore, while trying to satisfy the visitation requirements of some node $x \in X^L$ with $\mathcal{N}_x > 0$, the policy adopts an action selection scheme that maximizes the probability of visiting node x during a single traversal of the underlying graph $\mathcal{G}$. This action selection scheme can be computed from the LP of Theorem 2 and Equation 7 applied on a problem instance $\mathcal{E}_x$, that is obtained from the original problem instance, $\mathcal{E}$, by replacing the visitation requirement vector $\mathcal{N}$ with the unit vector $\mathbf{1}_x$. Letting V^*_x denote the optimal value of $\mathcal{E}_x$, it is also easy to see that the expected number of the graph traversals that are required by policy $\hat{\pi}$ in order to satisfy the

visitation requirements for node x, is equal to $\mathcal{N}_x \cdot V_x^*$. Hence, the value, $V^{\hat{\pi}}(s^0)$, of policy $\hat{\pi}$ – i.e., the expected traversals of the graph $\mathcal{G}$ in order to satisfy the entire set of the visitation requirements expressed by vector $\mathcal{N}$, under $\hat{\pi}$ – is equal to $\sum_{x \in X^L} \mathcal{N}_x \cdot V_x^*$. We summarize the above discussion in the following theorem.

Theorem 5 *Given a problem instance $\mathcal{E} = (X, \mathcal{A}, \mathcal{P}, \mathcal{N})$, an upper bound, $\hat{V}$, for its optimal value, V^*, is given by the expression*

$$\hat{V} = \sum_{x \in X^L} \mathcal{N}_x \cdot V_x^* \tag{41}$$

where the quantities V_x^*, for $x \in X^L$ with $\mathcal{N}_x > 0$, are the optimal values for the problem instances $\mathcal{E}_x$, that are obtained from the original problem instance, $\mathcal{E}$, by replacing the visitation requirement vector $\mathcal{N}$ with the unit vector $\mathbf{1}_x$.

Finally, we notice that the upper bound $\hat{V}$ provided by Theorem 5 is tight, since it is easy to construct problem instances in which the sequential strategy adopted by the underlying policy $\hat{\pi}$ is indeed an optimal strategy.

5 Conclusions

This paper introduced the problem of the optimal node visitation in acyclic stochastic digraphs and it provided its formal characterization as an SSP problem. It also established that the induced SSP problem possesses special structure that enables its solution through a spatial decomposition approach. This decomposing approach can alleviate the computational effort of computing an optimal policy, but it remains intractable for larger problem instances. Hence, an additional contribution of the presented work was the development of a randomized policy that can be deployed and executed with a computational cost that is polynomially related to the underlying problem size, and it is asymptotically optimal as the nodal visitation requirements grow uniformly to infinity. Finally, an additional outcome of the presented work was the derivation of a lower and an upper bound to the optimal value, V^*.

Future work will seek to rigorously resolve the computational complexity of this problem, and to derive bounds for V^* and policies of enhanced quality and performance; this second task will be especially important in the case that it is shown that the problem does not admit an optimal solution of polynomial complexity.

Appendix

Proof of Proposition 1

According to Definition 1, we need to show that there exists a stationary policy π such that in the Markov chain induced by π, every state $s \in S\backslash\{s^T\}$ is connected to the terminal state s^T with an action path of positive probability. We prove this result by a double induction where the outer induction runs on the size of the vector $\mathcal{N}^c$, defined by the l_1 norm, and the inner induction runs on the layer index of the node x. Hence, first consider a state $s = (x^0, \mathcal{N}^c)$ with $|\mathcal{N}^c| = 1$, and let the node $y \in X^L$ denote the unique terminal node with $\mathcal{N}^c_y > 0$. From the assumptions stated in the definition of the transition function $\mathcal{P}$, there exists an action path $\xi(y)$ that leads with positive probability from node x^0 to node y. Hence, there exists an action path of positive probability that leads from state s to the terminal state s^T. Next, consider a state $s = (x, \mathcal{N}^c)$ with $\mathcal{N}^c_y = 1$; $\mathcal{N}^c_z = 0,\ \forall z \in X^L\backslash\{y\}$, and a node $x \in X^L$. If $x = y$, state s^T is reachable from s by a single transition. If $x \neq y$, then, the unique action feasible at s leads deterministically from state s to state $(x^0, \mathcal{N}^c)$, which was shown to be connected to s^T by an action path of positive probability. Subsequently, consider a state $s = (x, \mathcal{N}^c)$ with $\mathcal{N}^c_y = 1$; $\mathcal{N}^c_z = 0,\ \forall z \in X^L\backslash\{y\}$, and a node $x \in X^{L-1}$. If this state belongs to the earlier constructed path leading from state $s = (x^0, \mathcal{N}^c)$ to s^T, then the sought action $\pi(x, \mathcal{N}^c)$ is selected to match the action suggested by that path. Otherwise, since by assumption $\mathcal{A}(x)$ is non-empty, pick any action $a \in \mathcal{A}(x)$. This action leads with positive probability to a number of states $s' = (x', \mathcal{N}^c)$ with $x' \in X^L$, for which we have already established paths connecting them to the terminal state s^T; selecting any of these states s' and the corresponding path will complete the argument for state s. Proceeding in a similar fashion with the subsequent layers $X^{L-2}, \ldots, X^1$, one can also establish a stationary policy π connecting every state $s = (x, \mathcal{N}^c)$ with $\mathcal{N}^c_y = 1$; $\mathcal{N}^c_z = 0,\ \forall z \in X^L\backslash\{y\}$, to the terminal state s^T with positive probability. Since the target node y was selected arbitrarily, and there is no communication among the sub-spaces defined by this selection, the above construction can be applied to every target node $y \in X^L$. Finally, repetition of the entire above argument will iteratively provide paths from states $s = (x, \mathcal{N}^c)$ with $|\mathcal{N}^c| = i$, for $i = 2, 3, \ldots, |\mathcal{N}|$, to the respective states $s' = (x^0, \mathcal{N}^{c'})$ with $|\mathcal{N}^{c'}| = i - 1$, and through them, to s^T.

Next,assume an improper policy π. Then, by definition, there is at least one state $s \in S\backslash\{s^T\}$ such that there is no action path of positive probability from state s to state s^T. Hence, when the process is initiated in state s and subsequently is operated under policy π, it will remain in the subspace $S\backslash\{s^T\}$ ad infinitum. Since, by the definition of $S\backslash\{s^T\}$, the process can undergo at most L transitions before incurring a positive cost, it follows that $V^\pi(s) = \infty$.

Proof of Proposition 2

To prove part (i) of Proposition 2, first notice that for any state $s = (x, \mathcal{N}^c)$ with $x \in X \backslash X^L$ and $\sum_{a \in \mathcal{A}(x)} \chi_a^* = 0$, $x \neq x_0$ (since, for $\mathcal{N} > 0$, $\sum_{a \in \mathcal{A}(x^0)} \chi_a^* > 0$). Next consider a state $s = (x, \mathcal{N}^c)$ with $x \in X \backslash X^L$ and $\sum_{a \in \mathcal{A}(x)} \chi_a^* = 0$. Then, Constraint 12 of the relaxing LP implies that the total flow entering node x under the optimal solution $\{\chi_a^* | a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x)\}$, is equal to zero. This last observation, when combined with the definition of the randomized policy π, imply that there is no action path of positive probability leading from state $s' = (x^0, \mathcal{N}^c)$ to state s. However, Equation 1 implies that state s' is the only state through which the underlying process can reach state s, when starting from the initial state $s^0 = (x^0, \mathcal{N})$, and therefore, state s is unreachable under policy π.

To prove part (ii) of Proposition 2, first notice that in the optimal solution $\{\chi_a^* | a \in \bigcup_{x \in X \backslash X^L} \mathcal{A}(x)\}$, there is a path of positive flow connecting the source node x^0 to any node $x \in X^L$ with $\mathcal{N}_x > 0$. In the operational context of the randomized policy π, each of these paths translates to an action path of positive probability leading from state $s = (x^0, \mathcal{N}^c)$ to the corresponding state $s' = (x, \mathcal{N}^c)$, for any $\mathcal{N}^c > 0$. Hence, in order to establish the required result, it is adequate to show that from any state $s'' \in S \backslash \{s^T\}$ that is reachable under policy π, we shall reach a state $s = (x^0, \mathcal{N}^c)$ or state s^T with probability one. The validity of this last statement follows immediately from (a) the result of part (i), established above, which guarantees the policy completeness at all the intermediately visited states, and (b) the transitional structure implied by Equation 1.

Proof of Lemma 1

We proceed by running an induction on the layer index, l. Clearly, $\mathcal{P}_{x^0} = 1$. Also, for $x \in X^1$, $\mathcal{P}_x$ can be expressed as $\sum_{a \in \mathcal{A}(x^0)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} \cdot p(x; a)$ and the theorem holds. Next assume that the theorem holds for all $x \in \bigcup_{i=1}^{l} X^i$ and let $x \in X^{l+1}$. Then we have

$$\mathcal{P}_x = \sum_{y \in \bigcup_{i=1}^{l} X^i} \mathcal{P}_y \sum_{a \in \mathcal{A}(y)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(y)} \chi_a^*} p(x;a) + \mathcal{P}_{x^0} \sum_{a \in \mathcal{A}(x^0)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} p(x;a) \tag{42}$$

$$= \sum_{y \in \bigcup_{i=1}^{l} X^i} \frac{\sum_{a:y \in \mathcal{S}(a)} p(y;a) \cdot \chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} \sum_{a \in \mathcal{A}(y)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(y)} \chi_a^*} p(x;a)$$

$$+ \sum_{a \in \mathcal{A}(x^0)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} p(x;a) \tag{43}$$

$$= \sum_{y \in \bigcup_{i=1}^{l} X^i} \frac{\sum_{a \in \mathcal{A}(y)} \chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} \sum_{a \in \mathcal{A}(y)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(y)} \chi_a^*} p(x;a) + \sum_{a \in \mathcal{A}(x^0)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} p(x;a) \tag{44}$$

$$= \frac{1}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} \sum_{y \in \bigcup_{i=1}^{l} X^i} \sum_{a \in A(y)} \chi_a^* \cdot p(x;a) + \sum_{a \in \mathcal{A}(x^0)} \frac{\chi_a^*}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} p(x;a) \tag{45}$$

$$= \frac{1}{\sum_{a \in \mathcal{A}(x^0)} \chi_a^*} \sum_{a:x \in S(a)} \chi_a^* \cdot p(x;a) \tag{46}$$

Notice that, as established in the proof of Proposition 2, the expression $\sum_{a \in \mathcal{A}(x^0)} \chi_a^*$ appearing in the denominator of the second term in the right-hand-side of Equation 42 will be strictly positive for well-defined problem instances, while any nodes y involved in the first term of the right-hand-side of Equation 42 that have $\sum_{a \in \mathcal{A}(y)} \chi_a^* = 0$, will also have $\mathcal{P}_y = 0$; hence, the right-hand-side of Equation 42 is well-defined. Furthermore, Equation 43 holds from the induction hypothesis, whereas Equation 44 holds from the equality constraints of the relaxing LP (c.f. Eq. 12). Thus, the induction is complete.

Acknowledgement

This work was partially supported by NSF grant DMI-MES-0318657.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[2] P. Billingsley. *Convergence of Probability Measures.* Wiley, N.Y., N.Y., 1968.

[3] P. Billingsley. *Probability and Measure (3rd edition).* Wiley, N.Y., N.Y., 1995.

[4] V. Chvátal. *Linear Programming.* W. H. Freeman & Co., N.Y., N.Y., 1983.

[5] S. A. Reveliotis and T. Bountourelis. Efficient pac learning for episodic tasks with acyclic state spaces. Technical Report (submitted to the Journal of Discrete Event Dynamic Systems), School of Industrial & Systems Eng., Georgia Tech, 2005.