Discovery of Optimal Solution Horizons in Non-Stationary Markov Decision Processes with Unbounded Rewards

Neustroev, Greg; de Weerdt, Mathijs; Verzijlbergh, Remco

Publication date
2019

Document Version
Final published version

Published in
Proceedings of the 29th International Conference on Automated Planning and Scheduling, ICAPS 2019

Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS 2019)

Discovery of Optimal Solution Horizons in Non-Stationary Markov Decision Processes with Unbounded Rewards

Grigory Neustroev, Mathijs de Weerdt, Remco Verzijlbergh
Delft University of Technology
Delft, The Netherlands

Abstract

Infinite-horizon non-stationary Markov decision processes provide a general framework to model many real-life decision-making problems, e.g., planning equipment maintenance. Unfortunately, these problems are notoriously difficult to solve, due to their infinite dimensionality. Often, only the optimality of the *initial* action is of importance to the decision-maker: once it has been identified, the procedure can be repeated to generate a plan of arbitrary length. The optimal initial action can be identified by finding a time horizon so long that data beyond it has no effect on the initial decision. This horizon is known as a solution horizon and can be discovered by considering a series of truncations of the problem until a stopping rule guaranteeing initial decision optimality is satisfied. We present such a stopping rule for problems with unbounded rewards. Given a candidate policy, the rule uses a mathematical program that searches for other possibly optimal initial actions within the space of feasible truncations. If no better action can be found, the candidate action is deemed optimal. Our rule runs faster than comparable rules and discovers shorter, more efficient solution horizons.

1 Introduction

Probabilistic planning is a long-standing challenge (Littman and Younes 2004), which arises in many domains, including inventory management (Shin and Lee 2015), equipment replacement (Hopp and Nair 1991), and robot surveillance (Witwicki et al. 2013). Infinite-horizon discounted Markov decision processes (MDPs) are often employed to model such problems, but they rely on a crucial but sometimes unrealistic assumption: the problem's data must remain constant. In order to incorporate possible temporal changes in the decision-making process, non-stationary (or non-homogeneous) MDPs (NS-MDPs) have been introduced (Hopp, Bean, and Smith 1987; Ghate and Smith 2013).

Unfortunately, NS-MDPs are infinitely-dimensional optimization problems by nature. This means that standard solution methods (e.g., value iteration and policy iteration) require an infinite number of calculations. To overcome this computational hurdle, NS-MDPs have been solved using a finite-time version of the problem, known as a truncation. Ghate (2011) provides a broad survey of such methods.

A typical approach is to use a rolling-horizon procedure (Sethi and Sorger 1991). At each time step the original infinite-horizon problem is truncated to a chosen time horizon, known as a *study horizon*, the truncation is solved, and the first decision is made based on this solution. The process is then repeated whenever another decision has to be made. While this approach is computationally feasible, it can lead to sub-optimal decisions, as the truncation by definition considers a limited horizon. Thus, it is important to identify a study horizon that is guaranteed to give the same initial decision as the infinite-horizon problem. This horizon is known as a *solution horizon* (Bès and Sethi 1988).

Due to unpredictability of future data and reduced computation time for smaller truncations, the decision-maker is often interested in a solution horizon that is as short as possible. The standard procedure for discovering such a horizon is to construct a series of longer and longer truncations until a certain condition is met. This condition, called a *stopping rule*, must guarantee that the last considered horizon is a solution horizon.

Several stopping rules have been proposed in the literature (Hopp, Bean, and Smith 1987; Bès and Lasserre 1986; Hernández-Lerma and Lasserre 1988; Cheevaprawatdomrong et al. 2007). Almost all of them are based on the assumption that the model data are uniformly bounded, and explicitly use these bounds. While uniform bounds are easy to work with, they can be very loose (e.g., if there is a large spike in data at one time step), providing inaccurate estimates of future states of the model. Moreover, for some problems the boundedness assumption may not hold at all.

Therefore, the goal of this paper is to develop a method applicable to unbounded problems. As an example of such a problem, consider a long-term investment project, such as a university endowment fund. The endowment needs to be divided between several different assets, such as stocks and bonds. In order to make the allocation, the decision-maker needs to model future returns on these assets. One of the most commonly employed models of discrete-time financial time series is a geometric random walk. In this case the returns r_t of stocks are modeled as $r_t = r_{t-1} + \mu + \sigma\varepsilon_t$, where $\varepsilon_t \sim N(0,1)$ is a sequence of i.i.d. random variables. Given current return r_0, future return forecasts are given by $r_t \sim N(r_0 + \mu t, \sigma^2 t)$. They can not be uniformly bounded for two reasons. Firstly, there is a slow unconditional growth

$\mu > 0$, known as drift, attributed to inflation, transaction costs and general economic growth. This drift makes expected returns $\mathbb{E}[r_t] = r_0 + \mu t$ grow linearly, meaning that any return value will be eventually exceeded. Secondly, forecasts farther into future become less precise, which means that their variance $\mathbb{V}\mathrm{ar}[r_t] = \sigma^2 t$ also increases with time, even if there is no drift (i.e., if $\mu = 0$), again without any limit. As a result, reward forecasts can not be bounded by a constant. However, bounds changing over time may exist.

In this paper, we propose a new stopping rule for infinite-horizon discounted NS-MDPs with unbounded rewards. Our rule searches for possible alternative initial decisions among the feasible problem truncations; if no such decision exists, the initial decision is deemed optimal, and the current horizon is a solution horizon. We show how the stopping rule can be implemented and demonstrate that it is capable of finding shorter solution horizons than existing methods.

2 Previous Work

Chand, Hsu, and Sethi (2002) provided an exhaustive review of literature on horizon methods. It shows that the majority of research in this area focuses on deterministic problems: of more than 200 papers reviewed, less than a third used stochastic models, including MDPs.

The most common approach is to exploit the cost (or reward) properties of a particular problem, both for deterministic and stochastic models. Two most commonly used properties are convexity (Smith and Zhang 1998; Cheevaprawatdomrong and Smith 2004) and supermodularity (Nair 1995; Cheevaprawatdomrong et al. 2007). For example, Nair (1995) considered an investment problem under sequential technological change. The proposed method is based on the assumption that future technologies will generate higher revenues than the current ones. While this assumption is not restrictive in the particular setting, such monotonically improving environment may not exist for other problems.

In the context of MDPs, Bès and Lasserre (1986) proposed a rolling-horizon procedure and a stopping rule based on the reward differences. Their stopping rule is elegantly simple: an initial decision is deemed optimal if it outperforms all other possible decisions by a given threshold. This threshold is chosen so as to guarantee that no matter what policy is employed after the solution horizon, the difference is outweighed by the initial decision. This method was later extended to the case of MDPs with Borel state spaces (Hernández-Lerma and Lasserre 1988).

Ergodic (i.e., related to recurrence of states) properties of the underlying Markov chains may be used as a source of solution horizons as well. For example, Hopp (1989) suggested the following stopping rule. For a given study horizon, approximate all of the future discounted rewards with some constants, known as salvage values. If for all feasible salvage values the resulting problems result in the same optimal initial decision, that decision must be optimal to the original infinite-horizon problem as well. Feasibility of the salvage values is established by bounding their spans using an ergodicity coefficient for discounting. The resulting space of possible salvage values forms a polyhedron, and linear programming can be used to solve the resulting problem (Bean, Hopp, and Duenyas 1992).

Another linear-programming based method for solving NS-MDPs was proposed by Ghate and Smith (2013). Even though their method addresses a slightly different problem and thus does not involve stopping rules, it provides some useful insights on the linear-programming formulations of NS-MDPs. Their results were later extended to the case of unbounded rewards in context of a more general class of countable-state MDPs (Lee et al. 2017).

As already mentioned, virtually all of the stopping rules require uniform bounds on the rewards (or their spans). The unbounded case remains relatively untreated. Cheevaprawatdomrong et al. (2007) provided a possible remedy, but necessarily introduced a different set of assumptions. To address this gap, we propose a modification of Hopp's stopping rule based on the results of Puterman (1994) and Lee et al. (2017) for countable-state MDPs and implement it using the linear programming method of Bean, Hopp, and Duenyas (1992). This modification is based on varying bounds instead of uniform ones, which allows us to construct substantially smaller spaces of possible salvage values, resulting in better and faster solutions.

3 NS-MDPs with Unbounded Rewards

This section formally introduces the problem of finding an optimal policy in an infinite-horizon discounted NS-MDP with unbounded rewards and shows how this problem can be reduced from a countably-infinite optimization problem to a finite one.

First, we introduce uniformly bounded NS-MDPs and define some properties of their optimal values. Next, we consider the case of unbounded rewards and show how it can be treated in a similar manner. Finally, we introduce an integer-programming formulation of the problem with unbounded rewards and its approximation, known as truncation.

The content of this section is based on existing work by Puterman (1994) and Lee et al. (2017) for countable-state MDPs. We introduced a simplified matrix-based notation and translated their results to the case of NS-MDPs, which is the main contribution of this section.

Preliminaries

An *infinite-horizon NS-MDP* is an MDP in which both the state space $\mathbb{S}$ and the action space $\mathbb{A}$ are discrete and finite, and the transition function $p_t : \mathbb{S}\times\mathbb{A} \to \mathbb{S}$ and the immediate reward function $r_t : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ are allowed to change over time. Without loss of generality we set $\mathbb{S} \triangleq \{1, 2, \ldots, S\}$ and $\mathbb{A} \triangleq \{1, 2, \ldots, A\}$.

Definition 1. A function $\pi : \mathbb{S} \times \mathbb{N}_0 \to \mathbb{A}$ is called a (deterministic Markov) *decision rule*. For each state-time tuple (s, t) it gives an action a. A sequence of decision rules $\pi \triangleq \{\pi_t\}_{t\in\mathbb{N}_0}$ is called a (deterministic Markov) *policy*. We denote the space of all policies as $\mathbb{D}$.

We denote the rewards and transitions under policy $\pi \in \mathbb{D}$ as $r_t^\pi(s)$ and $p_t^\pi(s'|s)$ respectively, that is,

$$r_t^\pi(s) \triangleq r_t\big(s, \pi_t(s)\big), \quad p_t^\pi(s'|s) \triangleq p_t\big(s'|s, \pi_t(s)\big);$$

we use $p_t^{\pi,j}(s'|s)$ for the j-step transition probability (i.e., the probability to reach state s' at time step $t+j$ by following policy π starting in state s at time step t).

At each time step $t \in \mathbb{N}_0$ the decision-maker observes the state s_t and chooses an action a_t according to a decision rule π_t of the chosen policy $\pi \in \mathbb{D}$. Given a state s_t at time t, each policy has a value $v_t^\pi(s)$ equal to the expected total γ-discounted reward:

$$v_t^\pi(s_t) \triangleq \mathbb{E}^\pi \sum_{\tau=t}^{\infty} \gamma^{t-\tau} r_t(s_t, a_t), \tag{1}$$

where $0 < \gamma < 1$. Assuming that the sum in equation (1) is well-defined for all policies, the decision-maker's goal is to find a policy π^* with the maximum value $v_0^*(s_0) = \max_\pi v_0^\pi(s_0)$, called an *optimal policy*.

Any NS-MDP can be rewritten as a stationary countable-state MDP (CS-MDP) by augmenting the state space $\mathbb{S}$ with time space $\mathbb{N}_0$. The new state space is given by $\mathbb{X} \triangleq \mathbb{S} \times \mathbb{N}_0$ and the transition probabilities p' and the rewards r' for all $x = (s,t)$ and $x' = (s',t')$ are equal to

$$p'\big(x'|x,a\big) \triangleq \delta_{t+1,t'} p_t(s'|s,a), \tag{2}$$

$$r'\big(x'|x,a\big) \triangleq \delta_{t+1,t'} r_t(s'|s,a), \tag{3}$$

where $\delta_{i,j} \triangleq \mathbb{I}_{\{i=j\}}$ denotes the Kronecker delta. Thus, properties of CS-MDPs can be translated back to NS-MDPs. In the remainder of this section we present some of these properties, translating them back to the original NS-MDP formulation of the problem using equations (2) and (3).

Let $\mathbb{V}$ be the space of all functions $v : \mathbb{X} \to \mathbb{R}$ with finite supremum-norm $\|v\| = \sup_{(s,t)\in\mathbb{X}} \big|v_t(s)\big|$.

Definition 2. An operator $\mathcal{L} : \mathbb{V} \to \mathbb{V}$ is called a *k-stage contraction* on $\mathbb{V}$ if there exists a constant $0 \le \lambda_{\mathcal{L},k} < 1$ such that for any $v' \in \mathbb{V}$ and $v'' \in \mathbb{V}$

$$\|\mathcal{L}^k v' - \mathcal{L}^k v''\| \le \lambda_{\mathcal{L},k} \|v' - v''\|.$$

If $k = 1$ then $\mathcal{L}$ is simply called a contraction.

Definition 3. A function $v \in \mathbb{V}$ is called a *fixed point* of an operator $\mathcal{L} : \mathbb{V} \to \mathbb{V}$ if $\mathcal{L}v = v$.

Any contraction and multi-stage contraction has a unique fixed point (Puterman 1994). When the rewards $r_t(s,a)$ are uniformly bounded, values $v_t^\pi(s)$ and $v_t^*(s)$ belong to $\mathbb{V}$ and are equal to the fixed points of operators $\mathcal{L}^\pi$ and $\mathcal{L}^*$ respectively, where

$$\mathcal{L}^\pi v_t(s) \triangleq r_t^\pi(s) + \gamma \sum_{s'\in\mathbb{S}} p_t^\pi(s'|s) v_{t+1}(s'), \tag{4}$$

$$\begin{aligned}\mathcal{L}^* v_t(s) &\triangleq \max_{a\in\mathbb{A}} \left\{ r_t(s,a) + \gamma \sum_{s'\in\mathbb{S}} p_t(s'|s,a) v_{t+1}(s') \right\} \\ &= \sup_{\pi\in\mathbb{D}} \mathcal{L}^\pi v_t(s). \end{aligned} \tag{5}$$

Both operators are contractions on $\mathbb{V}$.

To simplify the notation, we introduce the following matrices and vectors. All of these are understood as the values of the respective functions at a given time step t. For example, equation (4) can be written as

$$\mathcal{L}^\pi \mathbf{v}_t \triangleq \mathbf{r}_t^\pi + \gamma \mathbf{P}_t^\pi \mathbf{v}_{t+1},$$

where $\mathbf{r}_t^\pi$ are the immediate rewards and $\mathbf{P}_t^\pi$ are the one-step transition matrices under policy $\pi \in \mathbb{D}$, and $\mathbf{v}_t = [v_t(s)]_{s\in\mathbb{S}}$ are values of function v.

Using one-step transition matrices $\mathbf{P}_t^\pi$ we define j-step transition matrices $\mathbf{P}_t^{\pi,j} \triangleq \prod_{i=0}^{j-1} \mathbf{P}_{t+i}^\pi$. Additionally, we denote $\mathbf{P}_t^{\pi,0} \triangleq \mathbf{I}$, where $\mathbf{I}$ is an $S \times S$ identity matrix.

We denote the column-vector of expected immediate rewards for action a as $\mathbf{r}_{a,t} \triangleq [r_t(s,a)]_{s\in\mathbb{S}}$ and the vector of all expected immediate rewards as

$$\mathbf{r}_t \triangleq \left[\mathbf{r}_{1,t}^\top, \mathbf{r}_{2,t}^\top, \ldots, \mathbf{r}_{A,t}^\top\right]^\top.$$

Similarly, we define $S \times S$ transition matrices $\mathbf{P}_{a,t}$ as

$$\mathbf{P}_{a,t} \triangleq \big[p_t(j|i,a)\big]_{i,j=1}^{S}.$$

Let $\mathbf{P}_t$ be an $(S \cdot A) \times S$ matrix produced by stacking $\mathbf{P}_{a,t}$ for all actions a in a block-column:

$$\mathbf{P}_t \triangleq \left[\mathbf{P}_{1,t}^\top, \mathbf{P}_{2,t}^\top, \ldots, \mathbf{P}_{A,t}^\top\right]^\top.$$

Finally, we use $\mathbf{0} \triangleq [0,\ldots,0]^\top$, $\mathbf{i} \triangleq [1,\ldots,1]^\top$ to denote column-vectors of zeros and ones. Matrix $\mathbf{N} \triangleq [\mathbf{I},\ldots,\mathbf{I}]^\top$ denotes a block-column of A identity matrices $\mathbf{I}$; it has the same dimensions as matrices $\mathbf{P}_t$.

Unbounded Rewards

As mentioned, we are interested in the case of unbounded rewards, that is, when there exists no constant R such that $|r_t(s,a)| \le R$. In this case the policy value function $v^\pi(s)$ may not belong to $\mathbb{V}$, so equations (4) and (5) do not define contractions on $\mathbb{V}$. To address this issue, we assume that the reward can be bounded by a function $w : \mathbb{X} \to \mathbb{R}$ which changes over time, possibly unboundedly, but still guarantees that the values of different policies given by equation (1) are well-defined (Puterman 1994). This assumption is paramount to the results of this paper and is formalized as follows.

Assumption 1. *There exists a function $w : \mathbb{X} \to \mathbb{R}$ such that $\inf_{(s,t)\in\mathbb{X}} w_t(s) > 0$. Moreover, there exist constants $\kappa \ge 0$, $0 \le \lambda < 1$, and $J \in \mathbb{N}$ such that*

$$|\mathbf{r}_t| \le \mathbf{N}\mathbf{w}_t, \tag{6}$$

$$\mathbf{P}_t \mathbf{w}_{t+1} \le \kappa \mathbf{N} \mathbf{w}_t, \text{ and} \tag{7}$$

$$\gamma^J \mathbf{P}_t^{\pi,J} \mathbf{w}_{t+J} \le \lambda \mathbf{w}_t \tag{8}$$

for all $t \in \mathbb{N}_0$, $s \in \mathbb{S}$, $a \in \mathbb{A}$, and $\pi \in \mathbb{D}$.

In some cases an unbounded problem can be transformed into a bounded one. For example, if $\|w\| < \infty$, we can choose $R = \|w\|$ as a uniform bound. Alternatively, we can transform the problem as follows. Let $\tilde{\mathbb{S}} \triangleq \mathbb{S} \cup \{0\}$ and define

$$\tilde{r}_t(s,a) \triangleq \begin{cases} w_t^{-1}(s) \cdot r_t(s,a), & s \in \mathbb{S}, \\ 0, & s = 0, \end{cases} \tag{9}$$

$$\tilde{p}_t(s'|s,a) \triangleq \begin{cases} \frac{p_t(s'|s,a)\cdot w_{t+1}(s')}{\kappa w_t(s)}, & s \in \mathbb{S}, s' \in \mathbb{S}, \\ 1 - \sum\limits_{\tilde{s}\in\mathbb{S}} \frac{p_t(\tilde{s}|s,a)\cdot w_{t+1}(\tilde{s})}{\kappa w_t(s)}, & s \in \mathbb{S}, s' = 0, \\ 1, & s = s' = 0, \end{cases}$$

$$\tilde{\gamma} \triangleq \gamma\kappa.$$

Equation (7) guarantees that the probabilities in the transformed problem are less than one, and the absorbing state 0 is added so that they add up to one. The new problem is bounded by $R = 1$, and it is easy to check that its solution is equivalent to the solution of the original problem. Unfortunately, this method is only applicable if $\kappa < \gamma^{-1}$, otherwise the new discounting factor $\tilde{\gamma}$ is larger than one.

The existence of such function w guarantees that the values $\mathbf{v}_t^\pi$ of any policy $\pi \in \mathbb{D}$ are bounded by

$$|\mathbf{v}_t^\pi| \le L\mathbf{w}_t, \text{ and} \tag{10}$$
$$L \triangleq \begin{cases} \frac{J}{1-\lambda}, & \gamma\kappa = 1, \\ \frac{1}{1-\lambda} \cdot \frac{1-(\gamma\kappa)^J}{1-\gamma\kappa}, & \gamma\kappa \neq 1. \end{cases}$$

Consequently v^π may not belong to $\mathbb{V}$. However, we can define a different space $\mathbb{V}_w$ so that $v^\pi \in \mathbb{V}_w$.

Definition 4. The *w-weighted supremum norm* $\|\cdot\|_w$ of a function $v : \mathbb{X} \to \mathbb{R}$ is a norm given by

$$\|v\|_w \triangleq \sup_{(s,t)\in\mathbb{X}} w_t^{-1}(s) \cdot \big|v_t(s)\big|.$$

The space of functions with finite w-norm is denoted as $\mathbb{V}_w$.

Obviously, $v^\pi \in \mathbb{V}_w$ for any $\pi \in \mathbb{D}$, and operators $\mathcal{L}^\pi$ and $\mathcal{L}^*$ are J-step contractions on $\mathbb{V}_w$ with fixed points equal to $v_t^\pi(s)$ and $v_t^*(s)$ respectively (Puterman 1994).

Linear Programming Formulation

One of the possible approaches to find optimal policies in NS-MDPs uses linear programming based formulation presented in this subsection. Under Assumption 1 any NS-MDP can be represented by the following pair of countably-infinite linear programs (CILPs) (Lee et al. 2017):

$$\min_v \quad g(v) = \sum_{t=0}^{\infty} \mathbf{b}_t^\top \mathbf{v}_t \tag{P}$$
$$\text{s.t.} \quad \mathbf{N}\mathbf{v}_t - \gamma\mathbf{P}_t\mathbf{v}_{t+1} \ge \mathbf{r}_t, \quad \forall t \in \mathbb{N}_0, \tag{P.1}$$
$$v \in \mathbb{V}_w;$$

$$\max_y \quad f(y) = \sum_{t=0}^{\infty} \mathbf{r}_t^\top \mathbf{y}_t \tag{D}$$
$$\text{s.t.} \quad \mathbf{N}^\top\mathbf{y}_0 = \mathbf{b}_0,$$
$$\mathbf{N}^\top\mathbf{y}_{t+1} - \gamma\mathbf{P}_t^\top\mathbf{y}_t = \mathbf{b}_{t+1}, \quad \forall t \in \mathbb{N}_0,$$
$$\mathbf{y}_t \ge \mathbf{0}, \quad \forall t \in \mathbb{N}_0,$$
$$y \in L^1(\mathbb{X} \times \mathbb{A}),$$

where $L^1(\mathbb{Y})$ is the space of absolutely summable functions on $\mathbb{Y}$. This formulation is similar to known linear programming formulations for stationary MDPs (Puterman 1994), and for uniformly bounded NS-MDPs (Ross, Birnbaum, and Lukacs 1983).

The intuition behind the primal program (P) is as follows. Any $v \in \mathbb{V}_w$ satisfying the constraints (P.1) is known to be an upper bound for v^*. Minimization allows us to find a bound that is as tight as possible. v^* satisfies the constraints (P.1), so it is feasible for (P) and for any coefficients $\mathbf{b}_t > \mathbf{0}$ the program (P) achieves the minimum v^*, if the sum in $g(v)$ converges. This can be ensured by choosing $\mathbf{b}_t$ so that $\sum_{t=0}^{\infty} \mathbf{b}_t^\top \mathbf{w}_t < \infty$. In our case, coefficients $\mathbf{b}_t = \gamma^t\mathbf{i}$ satisfy this assumption, but other values may be chosen.

The dual problem (D) for the unbounded rewards case was derived by Lee et al. (2017). The pair of CILPs (P) and (D) exhibit strong duality, resulting in the following properties.

Lemma 1 (Theorems 3 and 4 of Lee et al. (2017))**.** *There exists a feasible solution to* (D) *such that for all $t \in \mathbb{N}_0$ and $s \in \mathbb{S}$ there exists exactly one a for which $y_t(s,a) > 0$ and $y_t(s,a') = 0$ for all $a' \neq a$. Moreover, the policy π that uses these actions is an optimal policy.*

Definition 5. The slack in (P.1) is called the *reduced cost of state-action pair (s,a) at time t*. We denote it as $n_t(s,a)$. Vectorized reduced costs $\mathbf{n}_t$ can be expressed as

$$\mathbf{n}_t \triangleq \mathbf{r}_t - \mathbf{N}\mathbf{v}_t + \gamma\mathbf{P}_t\mathbf{v}_{t+1}. \tag{11}$$

Reduced cost $n_t(s,a)$ is also known as an advantage of action a: it represents the benefit of taking action a over the optimal action. Therefore, it is always non-positive. Moreover, there exists a useful lower bound, as shown by the following lemma.

Lemma 2. $-\mathbf{h}_t \le \mathbf{n}_t \le \mathbf{0}$ *where* $\mathbf{h}_t \triangleq (L+\gamma\kappa L+1)\mathbf{N}\mathbf{w}_t$.

Proof. The upper bound follows from the constraints (P.1). The lower bound is derived using equations (6), (7), and (10), and (11):

$$\begin{aligned}\mathbf{n}_t &= \mathbf{r}_t - \mathbf{N}\mathbf{v}_t + \gamma\mathbf{P}_t\mathbf{v}_{t+1} \\ &\ge -\mathbf{N}\mathbf{w}_t - L\mathbf{N}\mathbf{w}_t - \gamma\kappa L\mathbf{N}\mathbf{w}_t = -\mathbf{h}_t. \end{aligned}$$ □

Lemma 3 (Theorems 2, 5 and 6 of Lee et al. (2017))**.** *Programs* (P) *and* (D) *are strongly dual (Lee et al. 2017), and the following complementary slackness conditions hold due to Lemma 1:*

$$y_t(s,a) \cdot n_t(s,a) = 0, \quad \forall t \in \mathbb{N}_0, s \in \mathbb{S}, a \in \mathbb{A}.$$

Problem Truncation

The CILP formulation is useful for analyzing mathematical properties of NS-MDPs, but cannot be solved directly, as it requires infinite computations: to find $\mathbf{v}_0$ one needs to know $\mathbf{v}_1$, which in turn requires $\mathbf{v}_2$, and so on. On the other hand, if at least one of the future value vectors $\mathbf{v}_{T+1}$ is actually known, all of the previous values $\mathbf{v}_0, \ldots, \mathbf{v}_T$ can be computed in finite time. Thus, one of the ways to reduce the problem is to replace future values $\mathbf{v}_{T+1}$ with some approximation $\mathbf{z}$. Even if the approximation is bad, the fact that $\mathcal{L}^\pi$ is a contraction means that each time it is applied to find a preceding value vector, the resulting values get closer to the fixed point (i.e., the true values). When T is sufficiently large, the initial values of the truncation will be close to those of the original problem.

Definition 6. A T-*truncation* of the problem (P) with *salvage vector* $\mathbf{z}$ is the problem given by

$$\begin{aligned} \min_{\mathbf{v}_0,\ldots,\mathbf{v}_T} \quad & g(\mathbf{v}_0,\ldots,\mathbf{v}_T) = \sum_{t=0}^{T} \mathbf{b}_t^\top \mathbf{v}_t \qquad \text{(P2)} \\ \text{s.t.} \quad & \mathbf{N}\mathbf{v}_t - \gamma \mathbf{P}\mathbf{v}_{t+1} \geq \mathbf{r}_t, \quad 0 \leq t < T, \\ & \mathbf{N}\mathbf{v}_T - \gamma \mathbf{P}\mathbf{z} \geq \mathbf{r}_T. \end{aligned}$$

The definition of truncation involves only one salvage vector $\mathbf{z}$. However, if we want to consider truncations of different lengths, we may want to use different salvage vectors. Therefore, instead of a salvage vector $\mathbf{z}$, we introduce a *salvage function* u. If $u \in \mathbb{V}_w$ the solutions of these truncations will be feasible solutions of the original problem (P).

Given such a function u, we obtain a *series of truncations* with different salvage vectors $\mathbf{u}_{T+1}$ at different study horizons T. Just like with other functions, we use $\mathbf{u}_t$ as a short-hand notation for all of the values of u at time t.

Additionally, we define operators $\mathcal{L}^{\pi}_{u,T} : \mathbb{V}_w \to \mathbb{V}_w$ and $\mathcal{L}^{*}_{u,T} : \mathbb{V}_w \to \mathbb{V}_w$ as follows:

$$\mathcal{L}^{\pi}_{u,T}\mathbf{v}_t \triangleq \begin{cases} \mathbf{r}^{\pi}_t + \gamma \mathbf{P}^{\pi}_t \mathbf{v}_{t+1}, & t < T, \\ \mathbf{r}^{\pi}_t + \gamma \mathbf{P}^{\pi}_t \mathbf{u}_{t+1}, & t = T, \\ \mathbf{u}_t, & t > T. \end{cases}$$

$$\mathcal{L}^{*}_{u,T}\mathbf{v}_t \triangleq \sup_{\pi \in \mathbb{D}} \mathcal{L}^{\pi}\mathbf{v}_t.$$

Both operators are multi-stage contractions, therefore, they have unique fixed points (Puterman 1994). We denote these points as $v^{\pi}_{u,T}$ and $v^{*}_{u,T}$, and their values at time t as $\mathbf{v}^{\pi}_{t,u,T}$ and $\mathbf{v}^{*}_{t,u,T}$. By properly choosing a salvage function u, we can obtain convergent upper or lower bounds on $\mathbf{v}^{*}_t$ using the following lemma.

Lemma 4 (Corollary 6.10.10 of Puterman (1994))**.** *If there exist functions u^- and u^+ in $\mathbb{V}_w$ such that $\mathcal{L}^{\pi}_{u,T}u^- \geq u^-$ and $\mathcal{L}^{\pi}_{u,T}u^+ \leq u^+$ for all $\pi \in \mathbb{D}$, then*

$$\mathbf{v}^{*}_{t,u^-,T} \leq \mathbf{v}^{*}_{t,u^-,T+1} \leq \mathbf{v}^{*}_{t} \leq \mathbf{v}^{*}_{t,u^+,T+1} \leq \mathbf{v}^{*}_{t,u^+,T}.$$

Definition 7. Functions u^- and u^+ of Lemma 4 are called lower and upper *value bounding functions*, and the values $v^{*}_{t,u^-,T}(s)$ and $v^{*}_{t,u^+,T}(s)$ are called *lower and upper approximations* respectively.

4 The Stopping Rule

Section 3 shows that NS-MDPs can be represented by CILPs. Even though these representations cannot be solved with finite computations, they can be approximated by truncations. As an approximation, a truncation may result in a solution with an immediate decision π_0 that is different from the optimal immediate decision of the original NS-MDP. Therefore, we are interested in a method that allows us to check optimality of this decision without solving the CILP. In this section we design such a method for NS-MDPs with unbounded rewards.

We start by presenting a problem formulation with variable salvage vector introduced by Hopp (1989) and demonstrate how it can be solved using a linear program of Bean, Hopp, and Duenyas (1992). Then we extend the results to NS-MDPs with unbounded rewards by introducing different salvage spaces based on bounding functions instead of uniform bounds. Finally, we present a new algorithm for discovery of optimal solution horizons that employs our stopping rule and exploits the fact that the Bellman operator of the unbounded problem is a multi-stage contraction.

Truncations with Variable Salvage Vector

Assume that for a given study horizon T and salvage function u we have solved a truncation and found the optimal initial action $\pi^{*}_{0,u,T}(s_0)$. We want to check if this action is equal to the optimal initial action $\pi^{*}_0(s_0)$ of the original problem.

Suppose that we know that values $\mathbf{v}^{*}_{T+1}$ belong to some sets $\mathbb{Z}_{T+1} \subseteq \mathbb{R}^S$. For example, if the values are non-negative and bounded from above by a constant R, $\mathbb{Z}_t$ can be S-dimensional cubes: $\mathbb{Z}_t = \{\mathbf{z} \mid \mathbf{0} \leq \mathbf{z} \leq R \cdot \mathbf{i}\}$. If all of the salvage vectors $\mathbf{z} \in \mathbb{Z}$ of a given subspace $\mathbb{Z} \subseteq \mathbb{R}^S$ result in T-truncations with the same optimal initial decision and optimal values $\mathbf{v}^{*}_{T+1}$ also belong to that set, then the original problem has the same optimal initial decision π^{*}_0 as the truncation. The following proposition formalizes this observation.

Proposition 1 (Generalized Hopp's stopping rule)**.** *Study horizon T is a solution horizon if the initial optimal action is the same for all $\mathbf{z} \in \mathbb{Z}_{T+1}$, where the sequence $\{\mathbb{Z}_t\}_{t\in\mathbb{N}_0}$ of subspaces $\mathbb{Z}_t \subseteq \mathbb{R}^S$ is chosen so that $\mathbf{v}^{*}_t \in \mathbb{Z}_t$.*

Proposition 1 was used in Hopp's stopping rule (Hopp 1989) for constant sequence $\mathbb{Z}_t = \mathbb{Z}$ based on the uniform bounds of the value vector spans. Given this stopping rule, solution horizons can be discovered by starting with a study horizon $T = 0$, checking the stopping rule, and incrementing T until the stopping rule is satisfied. In order for the rule to be of any practical use, we need to guarantee that this solution horizon discovery method terminates in finite time.

The salvage subspaces $\{\mathbb{Z}_t\}_{t\in\mathbb{N}_0}$ must be chosen so that the stopping rule is able to find a solution horizon. This condition can be satisfied due to the following lemma. If $\mathbb{Z} \subseteq \mathbb{V}_w$, where $\mathbb{Z}$ is the set of all salvage functions u providing salvage vectors $\mathbf{u}_t \in \mathbb{Z}_t$, then the stopping rule terminates due to the following proposition.

Proposition 2 (Generalized Lasserre and Bès 1984)**.** *If the optimal initial decision is unique, and $\mathbf{v}^{*}_0$ is finite, there exists a finite horizon T^* such that for any salvage function $u \in \mathbb{V}_w$ all T-truncations with $T \geq T^*$ have the same optimal initial decision.*

Note that the original statement of Proposition 2 only considers zero salvage function $u = 0$ and uniformly bounded rewards. Both conditions are only required to guarantee that the value functions are finite, and the proof remains the same if $u \in \mathbb{V}_w$ and $\|r\|_w \leq 1$. The only crucial assumption is that the action space $\mathbb{A}$ is finite.

Moreover, we need to ensure that the condition of the stopping rule can be checked in finite time. When $\mathbb{Z}_t$ are polyhedrons it can be done by solving a mixed integer linear program (Bean, Hopp, and Duenyas 1992).

First, we find the optimal initial decision rule $a_0 = \pi^*_{0,\mathbf{u},T}(s_0)$ for an arbitrary $\mathbf{u} \in \mathbb{Z}_{T+1}$. Then we allow the salvage vector $\mathbf{z}$ to vary within $\mathbb{Z}_{T+1}$ and seek a decision rule $\pi^*_{0,\mathbf{z},T}(s_0) \neq a_0$ by solving the following program:

$$\min_{\mathbf{z},\, \mathbf{v}_t,\, \tilde{\mathbf{y}}_t} \quad n_0 = \mathbf{j}_0^\top (\mathbf{r}_0 - \mathbf{N}\mathbf{v}_0 - \gamma \mathbf{P}_0 \mathbf{v}_1) \tag{P3}$$

s.t.

$$-\mathbf{h}_t \circ (\mathbf{i} - \tilde{\mathbf{y}}_t) \leq \mathbf{r}_t - \mathbf{N}\mathbf{v}_t + \gamma \mathbf{P}_t \mathbf{v}_{t+1} \leq \mathbf{0},\ 0 \leq t < T,$$
$$-\mathbf{h}_T \circ (\mathbf{i} - \tilde{\mathbf{y}}_T) \leq \mathbf{r}_T - \mathbf{N}\mathbf{v}_T + \gamma \mathbf{P}\mathbf{z} \leq \mathbf{0},$$
$$\mathbf{N}^\top \tilde{\mathbf{y}}_T = \mathbf{i},$$
$$\mathbf{j}_0^\top \tilde{\mathbf{y}}_0 = 0,$$
$$\mathbf{v}_t \in \mathbb{Z}_t, \qquad 0 \leq t < T,$$
$$\mathbf{z} \in \mathbb{Z}_{T+1},$$
$$\tilde{\mathbf{y}}_t \in \{0,1\}^{S \cdot A}, \qquad 0 \leq t < T,$$

where $\mathbf{j}_0$ is a vector of length $S \cdot A$, with all elements equal to zero except for the element corresponding to the state-action pair (s_0, a_0), which is equal to 1, so that $n_0 = n_0(s_0, a_0)$; constants $\mathbf{h}_t$ are defined in Lemma 2; $\mathbf{a} \circ \mathbf{b}$ stands for Hadamard (i.e., element-wise) product.

Program (P3) works as follows. By Lemma 3, if an optimal decision rule $a \neq a_0$ exists for some salvage vector $\mathbf{z}$, the reduced cost $n_0 \triangleq n_0(s_0, a_0)$ will be negative. We can check if n_0 can be made less than zero by minimizing it for all feasible values of $\mathbf{z}$ and variables of the primal-dual problem pair. By Lemmas 1–3, we only care for the sign of $\mathbf{y}_t$: if $y_t(s,a) > 0$, then $n_t(s,a) = 0$, and if $y_t(s,a) = 0$, $n_t(s,a) < 0$. Thus, we can replace $y_t(s,a)$ with binary variables $\tilde{y}_t(s,a) \triangleq \operatorname{sgn} y_t(s,a)$. The integer variables $\tilde{y}_t(s,a)$ must be added to ensure that the found solution is a feasible solution to the dual program.

The constraints of the program serve the following purposes. The expressions $\mathbf{r}_t - \mathbf{N}\mathbf{v}_t + \gamma \mathbf{P}_t \mathbf{v}_{t+1}$ in the first two constraints are equal to $\mathbf{n}_t$. Whenever $\tilde{y}_t(s,a) = 1$, the corresponding constraint becomes tight and ensures that $n_t(s,a) = 0$. When $\tilde{y}_t(s,a) = 0$, the left-hand side of the corresponding constraint becomes equal to $-h_t(s,a)$, and $n_t(s,a) > -h_t(s,a)$ always holds as per Lemma 2. Constraint $\mathbf{N}^\top \tilde{\mathbf{y}}_T = \mathbf{i}$ is equivalent to $\sum_{a \in \mathbb{A}} \tilde{y}_t(s,a) = 1$ for all $s \in \mathbb{S}$. It ensures that Lemma 1 holds.

Next, $\mathbf{j}_0^\top \tilde{\mathbf{y}}_0 = 0$ forces the program to search for policies with $\pi^*_{0,\mathbf{z},T}(s_0) \neq a_0$. This constraint is not necessary, and it can make the program infeasible if no actions other than a_0 are available for s_0, in which case a_0 is also optimal.

We add constraints $\mathbf{v}_t \in \mathbb{Z}_t$ to the formulation of Bean, Hopp, and Duenyas (1992), because in our case $\mathbf{v}^*_t$ is known to belong to $\mathbb{Z}_t$. These new constraints are not strictly necessary, but they help with speeding up computations by reducing the search space for variables $\mathbf{v}_t$.

We exclude constraints $\mathbf{v}_t \geq \mathbf{0}$ from the formulation of Bean, Hopp, and Duenyas (1992), as this assumption does not hold in our case. The non-negativity assumption was used to show that n_0 is zero only when a_0 is optimal, but Lemma 1 provides this result in our case.

Finally, we would like to note that it is not strictly necessary to solve the optimization problem: if at any iteration the solver finds a feasible solution with negative value of the objective function, it can proceed to the next study horizon.

In order to implement the program (P3) the salvage spaces $\mathbb{Z}_{T+1}$ need to be polyhedrons (i.e., we should be able to express them using sets of linear constraints). In the next subsection we provide such subspaces for the unbounded case.

Salvage Subspaces When Rewards Are Unbounded

To implement the program (P3) we need to be able to construct salvage subspaces $\mathbb{Z}_t$ so that they are polyhedrons and $\mathbf{v}^*_t \in \mathbb{Z}_t$.

Consider a case when value bounding functions u^+ and u^- of Definition 7 exist and are known. A sequence of spaces $\{\mathbb{Z}_t\}_{t \in \mathbb{N}_0}, \mathbb{Z}_t \subseteq \mathbb{R}^S$ where

$$\mathbb{Z}_t = \{\mathbf{z} \mid \mathbf{u}_t^- \leq \mathbf{z} \leq \mathbf{u}_t^+\}. \tag{12}$$

Due to Lemma 4, longer horizons will result in smaller ranges of possible optimal initial values, until eventually all of the truncations will agree in the optimal initial decision as per Proposition 2.

In the general case the only bounds on $\mathbf{v}^*_t$ are provided by (10) and the only salvage spaces we can use are

$$\mathbb{Z}_t = \{\mathbf{z} \mid -L\mathbf{w}_t \leq \mathbf{z} \leq L\mathbf{w}_t\}.$$

Unfortunately, these bounds are not value bounding functions in the sense of Lemma 4.

Nonetheless, functions $\pm Lw$ are the only information about the problem available in the most general case, so we want to establish similar properties for these functions. In order to do so, for any $j \in \mathbb{N}$ and $\pi \in \mathbb{D}$ we define an operator $\mathcal{L}^{\pi,j} : \mathbb{V}_w \to \mathbb{V}_w$ as

$$\mathcal{L}^{\pi,j}\mathbf{v}_t = (\mathcal{L}^\pi)^j \mathbf{v}_t = \sum_{i=0}^{j-1} \gamma^i \mathbf{P}_t^{\pi,i} \mathbf{r}_{t+i}^\pi + \gamma^j \mathbf{P}_t^{\pi,j} \mathbf{v}_{t+j},$$

and show the following property.

Lemma 5. *For all $\pi \in \mathbb{D}$, functions $u^\pm = \pm Lw$ satisfy $\mathcal{L}^{\pi,J} u^- \geq u^-$ and $\mathcal{L}^{\pi,J} u^+ \leq u^+$.*

Proof. We prove the statement for u^+; the proof for u^- is identical. Note that $L = \sum_{i=0}^{J-1} (\gamma\kappa)^i + \lambda L$ and recall equations (7) and (8). For all $t \in \mathbb{N}_0$

$$\begin{aligned}\mathcal{L}^{\pi,J}\mathbf{u}_t^+ &= \sum_{i=0}^{J-1} \gamma^i \mathbf{P}_t^{\pi,i} \mathbf{r}_{t+i}^\pi + \gamma^J L \mathbf{P}_t^{\pi,J} \mathbf{w}_{t+J} \\ &\leq \sum_{i=0}^{J-1} (\gamma\kappa)^i \mathbf{w}_t + \lambda L \mathbf{w}_t = L\mathbf{w}_t = \mathbf{u}_t^+. \qquad \square\end{aligned}$$

Lemma 4 uses the operators $\mathcal{L}^\pi$ to show that one-stage increments in study horizons lead to monotone convergence. In the unbounded case, Lemma 5 shows that similar properties hold if instead of looking one stage ahead, the decision-maker chooses J-stage increments in study horizons, as operators $\mathcal{L}^\pi$ are now J-stage contractions. This is a crucial property leveraged by our algorithm; it ensures that the space of possible initial values decreases with each iteration, and the algorithm converges monotonically.

The Algorithm

Summarizing the aforementioned results, we present Algorithm 1. It is guaranteed to terminate in a finite number of steps if the optimal policy is unique. Moreover, when better value bounding functions are known, they can be used instead of $\pm Lw$ to provide smaller salvage subspaces $\mathbb{Z}_t$, resulting in faster convergence.

Algorithm 1: Solution horizon discovery

Data: an NS-MDP with a bounding function w.
Result: an optimal initial action a_0^* and a solution horizon T.

Let $u^+ \leftarrow Lw$ and $u^- \leftarrow -Lw$;
for $N \leftarrow 1, 2, \ldots$ **do**
  $T \leftarrow N \cdot J - 1$;
  solve (P2) with any salvage vector $\mathbf{z} \in \mathbb{Z}_{T+1}$;
  solve (P3) with $\mathbb{Z}_t$ given by (12);
  if (P3) *is infeasible or* $n_0 = 0$ **then**
    $a_0^* \leftarrow \pi^*_{0,z,T}(s_0)$;
    $T^* \leftarrow T$;
    break
for $n \leftarrow 1, \ldots J-1$ **do**
  $T \leftarrow T^* - n$;
  solve (P3) with $\mathbb{Z}_t$ given by (12);
  if (P3) *is feasible and* $n_0 < 0$ **then**
    return $a_0^*, T+1$;
    break

The algorithm searches for a solution horizon by doing J-stage increments in study horizons. For each of these horizons it checks if all of the feasible truncations agree in the initial optimal decision. Once a solution horizon has been identified, the algorithm returns back in time, up to the previous considered study horizon. It does so in order to identify possible shorter solution horizons.

5 Experimental Results

To demonstrate the performance of our stopping rule, we implemented Algorithm 1 for the following problem, known as an equipment replacement problem (Bean, Hopp, and Duenyas 1992).

Consider a piece of equipment subject to deterioration. The state space $\mathbb{S} = \{1, \ldots, S\}$ represents the state of its decay, with 1 being "new". At each time step, the decision-maker chooses between two actions: "replace" (action 1) and "keep" (action 2).

Transition probabilities of the problem are given by

$$p_t(s'|s,1) = \begin{cases} 1, & s' = 1, \\ 0, & \text{otherwise;} \end{cases}$$

$$p_t(s'|s,2) = \begin{cases} 1-\psi, & s' = s < S, \\ \psi, & s' = s+1, s' < S, \\ 1, & s' = s = S, \\ 0, & \text{otherwise,} \end{cases}$$

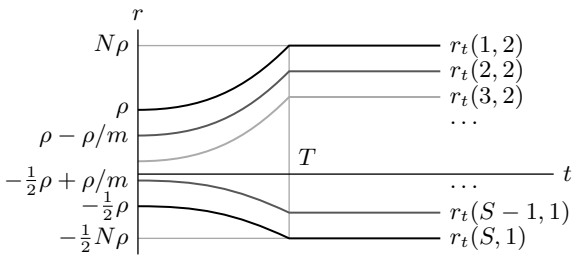

Figure 1: Rewards in the equipment replacement problem

where ψ is the deterioration probability. If the equipment is replaced, the state always changes to 1 (i.e., "new"). Otherwise, it either deteriorates to the next state (if there is one) with probability ψ, or remains the same state with probability $1 - \psi$.

In the first experiment we used the following rewards:

$$r_t(s,1) = \rho \cdot \left(-0.5 N^{\min\{t/T,1\}} + (S-s)/m\right);$$
$$r_t(s,2) = \rho \cdot \left(N^{\min\{t/T,1\}} - (s-1)/m\right).$$

Figure 1 outlines the general reward structure. When the equipment is kept, it generates revenue which depends on the state of deterioration and grows over time. If the equipment is new, the initial revenue $r_0(1,2)$ is equal to ρ and it grows exponentially (e.g., due to inflation). For each stage of deterioration the revenue decreases by ρ/m. When the equipment is replaced, it generates no revenue, and a replacement cost needs to be paid. The costs behave similarly to revenues, and the worse is the state of the equipment, the larger are the costs. We limit the data at time step T, when it becomes equal to $N \cdot \rho$ to add uniform bounds.

Function $w_t = \rho \cdot N^{\min\{t/T,1\}}$ satisfies Assumption 1 with $\kappa = N^{1/T}$. Assuming $T > -\log_\gamma N$, $\lambda = \gamma\kappa$ and $J = 1$, so functions $\pm Lw_t$ can be used as bounds for the state values. These are *loose bounds*, as they don't use any additional information. The following functions can be used as tighter bounding functions:

$$u_t^+ = \sum_{\tau=0}^{\infty} \gamma^\tau \max_{s,a} r_{t+\tau}(s,a) = \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}(1,2),$$
$$u_t^- = \sum_{\tau=0}^{\infty} \gamma^\tau \min_{s,a} r_{t+\tau}(s,a) = -\frac{1}{2} u_t^+.$$

These *tighter bounds* are easy to compute and result in smaller search spaces $\mathbb{Z}_t$, making the problem easier to solve. In practice, for a truly non-stationary problem such closed-form bounds will not be available, therefore they can be seen as a bound of what can be achieved without exploiting any additional information on the exact reward structure.

Stationarity of the rewards after the capping horizon T allowed us to find the exact solution of the problem. We started at time horizon T and solved the problem using value iteration, then used dynamic programming to obtain the initial optimal decision.

We compared our stopping rule for both choices of the bounding functions to Hopp's rule. Both methods were implemented in Python 3.6.4; `numpy` package was used to

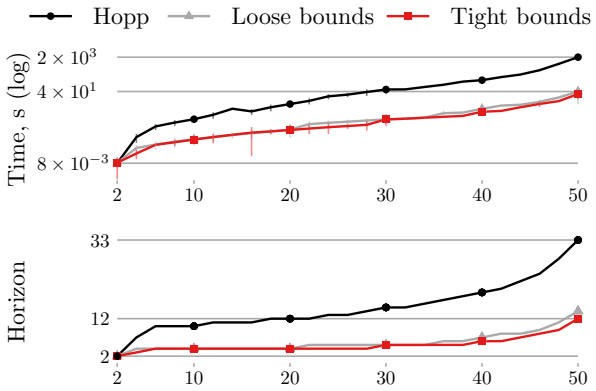

Figure 2: Performance with respect to the problem size S

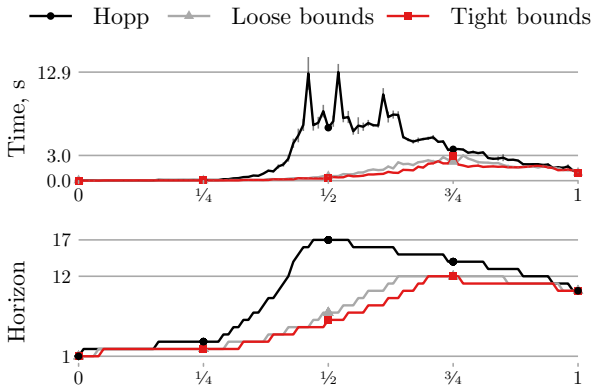

Figure 3: Performance with respect to deterioration rate ψ

work with vector data, and the linear programs were solved using Gurobi 8.0.1. The tests were performed on a computer with a 3.1 GHz Intel Core i5 two-core CPU and 8 GB of memory (MacBook Pro, 13-inch, 2017).

We ran the experiments for different combinations of parameters. For all of them, both stopping rules identified the optimal initial action correctly but discovered different solution horizons. In almost all of the experiments, our stopping rule was able to find a significantly shorter solution horizon. The following default values were used in all of the experiments, unless stated otherwise:

$$S = 10, \quad N = 10, \quad T = 10^3, \quad m = 45,$$
$$\gamma = 0.95, \quad \psi = 0.4, \quad \rho = 1, \quad s_0 = 1.$$

Figure 2 shows how the solution horizons and run-times scale with respect to the number of states S. Both algorithms need to look further into the future as the problem size grows, however, our stopping rule identifies significantly shorter solution horizons. Shorter horizons mean that less mixed-integer programs need to be solved, which substantially reduces the run-time.

Figure 3 presents the effect of the model uncertainty ψ on the algorithm. The largest difference in performance is exhibited when $\psi = 0.5$, that is, when the system's entropy is the largest.

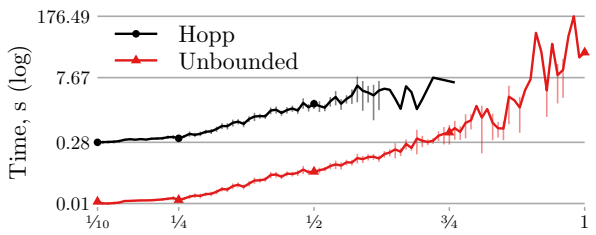

Figure 4: Performance with respect to spectral radius σ

In the second experiment, we set $S = 5$ and used the same transition matrices but different rewards. We randomly generated initial rewards $\mathbf{r}_{a,0}$ from the following sets

$$\mathbf{r}_{1,0} \in [-0.5N, 0)^S, \quad \mathbf{r}_{2,0} \in [0, N)^S.$$

Subsequent rewards were given by $\mathbf{r}_{a,t+1} = \mathbf{\Phi}_a \mathbf{r}_{a,t} = \mathbf{\Phi}_a^t \mathbf{r}_{a,0}$, where $\mathbf{\Phi}_a$ are tri-diagonal matrices with non-zero elements drawn from uniform distribution on $[-1, 1)$, and then scaled so that spectral radii σ_a of $\mathbf{\Phi}_a$ were less than one. The latter condition was added to ensure that the problem has a bounding function w. These spectral radii are similar to discounting factors for matrices, because they indicate the rate of growth of matrix power series; therefore for problems with $\sigma = \max\{\sigma_1, \sigma_2\} \geq 1$ the values $\mathbf{v}_t$ may not be well-defined.

The rewards of this problem are bounded by the function

$$w_t(s) = N \cdot \max\{\|\mathbf{\Phi}_1^t\|, \|\mathbf{\Phi}_2^t\|\}$$

with the following coefficients:

$$J = \min_{j \in \mathbb{N}_0} \{j : \|\gamma^j \mathbf{\Phi}_1^j\| < 1 \wedge \|\gamma^j \mathbf{\Phi}_2^j\| < 1\},$$

$$\kappa = \max\{\|\mathbf{\Phi}_1\|, \|\mathbf{\Phi}_2\|\}, \ \lambda = \max\{\|\gamma^J \mathbf{\Phi}_1^J\|, \|\gamma^J \mathbf{\Phi}_2^J\|\}.$$

Existence of J is guaranteed by the following property of spectral radii: $\sigma_a = \lim_{t \to \infty} \|\mathbf{\Phi}_a^t\|^{1/t}$. As a result, for any $\sigma < 1$ the norm $\|\mathbf{\Phi}_a^t\|$ becomes less than one eventually.

When $\gamma\kappa < 1$, the problem can be transformed into a bounded problem by using (9). In this case we are able to solve the problem using Hopp's stopping rule as well.

In this experiment the data was truly non-stationary, and it was impossible to compute value functions exactly. When Hopp's method was able to solve the problem, we knew that the action it identified was indeed optimal and used it as a benchmark for our method.

The results are presented in Figure 4. In all of the experiments our method was able to identify the optimal initial action. In these cases our method always returned the same horizon as Hopp's. This can be explained by the fact that the methods are similar: after the transformation is applied to the problem the salvage spaces $\mathbb{Z}_t$ become identical at all time steps, just like in Hopp's case.

Nevertheless, our method runs faster, as, on the one hand, it does not require the data transformation, and on the other hand, it uses large steps J when searching for the solution horizon, reducing the number of iterations by a factor of J. Moreover, it is applicable to a wider range of problems: for example, Hopp's stopping rule cannot be used in problems with a large spectral radius, as illustrated by Figure 4.

6 Conclusions

This paper presented a stopping rule for discovery of solution horizons in non-stationary Markov decision processes. The rule is applicable to problems with unbounded rewards and does not require any additional assumptions on the reward structure, such as convexity of rewards, making it applicable to a broad class of problems. An experimental study showed that our stopping rule was able to find better solution horizons more quickly even when the rewards can be uniformly bounded.

Future research directions include a formal proof of the algorithm's monotone convergence, and an extension to problems with countably-infinite base state spaces, as the problem is already countably-infinite in the time domain. Additionally, the rate of convergence might be improved by considering span-based bounds in combination with supremum-norm ones.

7 Acknowledgements

This research received funding from the Netherlands Organisation for Scientific Research (NWO).

References

Bean, J. C.; Hopp, W. J.; and Duenyas, I. 1992. A stopping rule for forecast horizons in nonhomogeneous Markov decision processes. *Operations Research* 40(6):1188–1199.

Bès, C., and Lasserre, J. 1986. An on-line procedure in discounted infinite-horizon stochastic optimal control. *Journal of Optimization Theory and Applications* 50(1):61–67.

Bès, C., and Sethi, S. P. 1988. Concepts of forecast and decision horizons: Applications to dynamic stochastic optimization problems. *Mathematics of Operations Research* 13(2):295–310.

Chand, S.; Hsu, V. N.; and Sethi, S. 2002. Forecast, solution, and rolling horizons in operations management problems: A classified bibliography. *Manufacturing & Service Operations Management* 4(1):25–43.

Cheevaprawatdomrong, T., and Smith, R. L. 2004. Infinite horizon production scheduling in time-varying systems under stochastic demand. *Operations Research* 52(1):105–115.

Cheevaprawatdomrong, T.; Schochetman, I. E.; Smith, R. L.; and Garcia, A. 2007. Solution and forecast horizons for infinite-horizon nonhomogeneous Markov decision processes. *Mathematics of Operations Research* 32(1):51–72.

Ghate, A., and Smith, R. L. 2013. A linear programming approach to nonstationary infinite-horizon Markov decision processes. *Operations Research* 61(2):413–425.

Ghate, A. 2011. Infinite horizon problems. *Wiley Encyclopedia of Operations Research and Management Science*.

Hernández-Lerma, O., and Lasserre, J. B. 1988. A forecast horizon and a stopping rule for general Markov decision processes. *Journal of Mathematical Analysis and Applications* 132(2):388–400.

Hopp, W. J., and Nair, S. K. 1991. Timing replacement decisions under discontinuous technological change. *Naval Research Logistics* 38(2):203–220.

Hopp, W. J.; Bean, J. C.; and Smith, R. L. 1987. A new optimality criterion for nonhomogeneous Markov decision processes. *Operations Research* 35(6):875–883.

Hopp, W. J. 1989. Identifying forecast horizons in nonhomogeneous Markov decision processes. *Operations Research* 37(2):339–343.

Lasserre, J., and Bès, C. 1984. Infinite horizon nonstationary stochastic optimal control problem: A planning horizon result. *IEEE Transactions on Automatic Control* 29(9):836–837.

Lee, I.; Epelman, M. A.; Romeijn, H. E.; and Smith, R. L. 2017. Simplex algorithm for countable-state discounted Markov decision processes. *Operations Research* 65(4):1029–1042.

Littman, M. L., and Younes, H. L. 2004. Introduction to the probabilistic planning track. In *Online Proceedings for The Probablistic Planning Track of IPC-04*.

Nair, S. K. 1995. Modeling strategic investment decisions under sequential technological change. *Management Science* 41(2):282–297.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley New York.

Ross, S. M.; Birnbaum, Z. W.; and Lukacs, E. 1983. *Introduction to Stochastic Dynamic Programming*. Academic Press.

Sethi, S., and Sorger, G. 1991. A theory of rolling horizon decision making. *Annals of Operations Research* 29(1):387–415.

Shin, J., and Lee, J. H. 2015. MDP formulation and solution algorithms for inventory management with multiple suppliers and supply and demand uncertainty. In Gernaey, K. V.; Huusom, J. K.; and Gani, R., eds., *12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering*, volume 37 of *Computer Aided Chemical Engineering*. Elsevier. 1907–1912.

Smith, R. L., and Zhang, R. Q. 1998. Infinite horizon production planning in time-varying systems with convex production and inventory costs. *Management Science* 44(9):1313–1320.

Witwicki, S.; Melo, F.; Capitán, J.; and Spaan, M. 2013. A flexible approach to modeling unpredictable events in MDPs. In *Twenty-Third International Conference on Automated Planning and Scheduling*, 260–268.