# Computational Approaches for Stochastic Shortest Path on Succinct MDPs

**Krishnendu Chatterjee**[1], **Hongfei Fu**[2,†], **Amir Goharshady**[1] and **Nastaran Okati**[3]
[1]IST Austria
[2]Shanghai Jiao Tong University
[3]Ferdowsi University of Mashhad
kchatterjee@ist.ac.at, fuhf@cs.sjtu.edu.cn, goharshady@ist.ac.at, nastaran.okati@mail.um.ac.ir

## Abstract

We consider the stochastic shortest path (SSP) problem for succinct Markov decision processes (MDPs), where the MDP consists of a set of variables, and a set of nondeterministic rules that update the variables. First, we show that several examples from the AI literature can be modeled as succinct MDPs. Then we present computational approaches for upper and lower bounds for the SSP problem: (a) for computing upper bounds, our method is polynomial-time in the implicit description of the MDP; (b) for lower bounds, we present a polynomial-time (in the size of the implicit description) reduction to quadratic programming. Our approach is applicable even to infinite-state MDPs. Finally, we present experimental results to demonstrate the effectiveness of our approach on several classical examples from the AI literature.[‡]

## 1 Introduction

*Markov decision processes.* Markov decision processes (MDPs) [Howard, 1960] are a standard mathematical model for sequential decision making, with a wide range of applications in artificial intelligence and beyond [Puterman, 1994; Filar and Vrieze, 1997; Bertsekas, 2005]. An MDP consists of a set of states, a finite set of actions (that represent the nondeterministic choices), and a probabilistic transition function that describes the transition probability over the next states, given the current state and action. One of the most classical optimization objectives in MDPs is the stochastic shortest path (SSP) problem, where the transitions of the MDP are labeled with rewards/costs, and the goal is to optimize the expected total rewards until a target set is reached.

*Curse of dimensionality.* In many typical applications, the computational analysis of MDPs suffers from the curse of dimensionality. The state space of the MDP is huge as it represents valuations to many variables that constitute the MDP. A well-studied approach for algorithmic analysis of large MDPs is to consider factored MDPs [Guestrin *et al.*, 2003; Delgado *et al.*, 2011], where the transition and reward function dependency can be factored only on few variables.

*Succinct MDPs.* In the spirit of factored MDPs, which aim to deal with the curse of dimensionality, we consider *succinct MDPs*, where the MDP is described implicitly by a set of variables, and a set of rules that describe how the variables are updated. The rules can be chosen non-deterministically from this set at every time step to update the variables, until a target set (of valuations) is reached. The rules and the target set represent the succinct description of the MDP. We consider the SSP problem for succinct MDPs.

*Our contributions.* Our main contributions are as follows:

1. First, we show that many examples from the AI literature (e.g., Gambler's Ruin, Robot Planning, and variants of Roulette) can be naturally modeled as succinct MDPs.
2. Second, we present mathematical and computational results for the SSP problem for succinct MDPs. For the SSP problem the $\sup$-value (resp. $\inf$-value) represents the expected shortest path value with supremum (resp., infimum) over all policies. We consider linear bounds for the SSP problem and our algorithmic bounds are as follows: (a) for the $\sup$-value (resp. $\inf$-value) we show that an upper (resp., lower) bound can be computed in polynomial time in the implicit description of the MDP; (b) for the $\sup$-value (resp. $\inf$-value) we show that a lower (resp., upper) bound can be computed by a polynomial-time (in the implicit description) reduction to quadratic programming. Our approach is as follows: we use results from probability theory to establish a mathematical approach to compute bounds for the SSP problem for succinct MDPs (Section 3), and reduce the mathematical approach to constraint solving to obtain a computational method (Section 4).
3. Finally, we present experimental results on several classical examples from the literature where our method computes tight bounds (i.e., lower and upper bounds are quite close) on the SSP problem extremely efficiently.

*Comparison with approaches for factored MDPs.* Some key advantages of our approach are the following. First, our approach gives a provably polynomial-time algorithm

† Corresponding Author

‡A full version of this paper, including details and formal proofs, is available at [Chatterjee *et al.*, 2018].

or polynomial-time reduction to quadratic programming in terms of the size of the implicit description of the MDP. Second, while algorithms for factored MDPs are typically suitable for finite-state MDPs, our method can handle MDPs with countable state space (e.g., when the domains of the variables are integers), or even uncountable state space (e.g., when the domains of the variables are reals). See Remark 1 for details.

## 2 Definitions and Examples

We define succinct Markov Decision Processes, the stochastic shortest path problem, and provide illustrative examples.

### 2.1 Succinct Markov Decision Processes

*Markov decision processes (MDPs).* MDPs are a standard mathematical model for sequential decision making. An MDP consists of a state space $S$ and a finite action space $A$, and given a state $s$ and an action $a$ permissible in $s$, there is a probability distribution function $P$ that describes the transition probability from the current state to the next states (i.e., $P(s' \mid s, a)$ is the transition probability from $s$ to $s'$ given action $a$). Moreover, there is a reward assigned to each state and action pair.

*Factored MDPs.* In many applications of MDPs, the state space of the MDP is high-dimensional and huge (i.e., the state space consists of valuations to many variables). For computational analysis it is often considered that the MDP can be factored, i.e., the transition and reward probabilities depend only on a small number of variables. For algorithmic analysis of finite-state factored MDPs see [Guestrin *et al.*, 2003].

*Succinct MDPs.* In this work, we consider succinct MDPs, which are related to factored MDPs. First, we present an informal description and then the details. A *succint MDP* is described by a set of variables, and a set of rules that update them. The update rule can be chosen non-deterministically or stochastically. Thus the MDP is described implicitly with the set of rules and a condition that describes the target set of states, and the MDP terminates when the target set is reached. We describe succinct MDPs as a special class of programs with a single while loop.

*Simple-while-loop succinct MDPs.* We consider succinct MDPs described by a simple while loop program. There are two types of variables, namely, *program* and *sampling* variables. Program variables are normal variables, while sampling variables are those whose values are sampled independently wrt some probability distribution. In general, both program and sampling variables can take integer, or even real values. The succinct MDP is described by a simple while loop of the form:

$$\textbf{while } \phi \textbf{ do } Q_1 \square \ldots \square Q_k \textbf{ od} \tag{1}$$

- $\phi$ is the *loop guard* defined as a single comparison between linear arithemtic expressions over program variables (e.g. $x \leq y + 1$);
- in the loop body, $Q_1, \ldots, Q_k$ are sequential compositions of assignment statements, grouped by the nondeterminism operator $\square$.

Every assignment statement has a program variable as its left-hand-side and a linear arithmetic expression over program and sampling variables as its right-hand-side. The operator $\square$ is the nondeterministic choice which means that the decision as to which $Q_i$ will be executed in the current loop iteration depends on a scheduler (or policy) that resolves nondeterminism. We first provide the formal syntax and then an example.

*Formal syntax of simple-while-loop programs.* A succinct MDP is specified by a simple-while-loop program equipped with probability distributions for sampling variables. We now formalize the intuitive description provided in Equation (1). Formally, a simple-while-loop program can be produced using the following grammar, where each `<pvar>` is chosen from a finite fixed set $X$ of program variables, each `<svar>` from a finite fixed set $R$ of sampling variables and each `<constant>` denotes a floating point number:

⟨*simple-while-loop-program*⟩ ::= '`while`' ⟨*guard*⟩ '`do`' '`{`' ⟨*nondet-block-list*⟩ '`}`' '`od`'

⟨*guard*⟩ ::= ⟨*linear-pvar-expr*⟩ ⟨*cmp*⟩ ⟨*linear-pvar-expr*⟩

⟨*linear-pvar-expr*⟩ := ⟨*constant*⟩ | ⟨*constant*⟩ '$\star$' ⟨*pvar*⟩
| ⟨*linear-pvar-expr*⟩ '+' ⟨*linear-pvar-expr*⟩

⟨*cmp*⟩ ::= '>=' | '>' | '<=' | '<'

⟨*nondet-block-list*⟩ ::= ⟨*block*⟩
| ⟨*block*⟩ $\square$ ⟨*nondet-block-list*⟩

⟨*block*⟩ ::= ⟨*assignment*⟩ | ⟨*assignment*⟩ ⟨*block*⟩

⟨*assignment*⟩ ::= ⟨*pvar*⟩ '`:=`' ⟨*linear-expr*⟩ '`;`'

⟨*linear-expr*⟩ ::= ⟨*constant*⟩ | ⟨*constant*⟩ '$\star$' ⟨*pvar*⟩
| ⟨*constant*⟩ '$\star$' ⟨*svar*⟩
| ⟨*linear-expr*⟩ '+' ⟨*linear-expr*⟩

**Example 1.** *Consider the following program*

$$\textbf{while } x \geq 1 \textbf{ do } x := x + r \square x := x - 1 \textbf{ od}$$

*where $x$ is a program variable and $r$ is a sampling variable that observes the two-point distribution $\mathbb{P}(r = -1) = \mathbb{P}(r = 1) = \frac{1}{2}$. Informally, the program performs either decrement or random increment/decrement on $x$ until its value is zero.*

*Informal description of the semantics.* Given a simple-while-loop program in the form (1), an MDP is obtained as follows: the state space $S$ consists of values assigned to program variables (i.e., *valuations* for program variables); the action space $A$ correponds to the non-deterministic choice between $Q_1, \ldots, Q_k$; and the transition function $P$ depends on the sampling of the sampling variables, which given the current valuation for program variables probabilistically updates the valuation. The assignments are linear functions, and the loop guard $\phi$ describes the target states as those which do not satisfy $\phi$. Given the above components, the notion of a policy (or scheduler) $\sigma$ that resolves the non-deterministic choices,

and that of the probability space $\mathbb{P}^\sigma$ given a policy is standard [Puterman, 1994]. For a more formal treatment of the semantics, see [Chatterjee *et al.*, 2018].

**Remark 1** (Simple-while-loop MDPs and Factored MDPs)**.** *The principle behind factored MDPs and simple-while-loop succinct MDPs is similar. Both aim to consider high-dimensional and large MDPs described implicitly in a succinct way. In factored MDPs the transitions and reward functions can be factored based on small sets of variables, but the dependency on the variables in these sets can be arbitrary. In contrast, in simple-while-loop succinct MDPs, we only allow linear arithmetic expressions as guards and assignments. Moreover, our MDPs do not allow nesting of while loops. The goal of our work is to consider linear upper and lower bounds, and nesting of linear loops can result in super-linear bounds. Hence, we do not consider nesting of loops. Therefore, simple-while-loop succinct MDPs are a special class of factored MDPs.*

*For algorithmic approaches with theoretical guarantees on computational complexity, the analysis of factored MDPs has typically been restricted to finite-state MDPs. However, we will present solutions for simple-while-loop programs, where the variables can take integer or real values, and thus the underlying state space is infinite or even uncountable. Thus the class of simple-while-loop MDPs consists of large finite-state MDPs (when the variables are bounded); countable state MDPs (variables are integer); and even uncountable state MDPs (variables are real-valued). Moreover, our algorithmic approaches provide computational complexity guarantee on the input size of the implicit representation of the MDP. Note that for finite-state MDPs, the implicit representation can be exponentially more compact than the explicit MDP. For example, $n$ boolean variables lead to a state-space of size $2^n$.*

In the sequel we consider MDPs obtained from simple-while-loop programs and, for brevity, call them succinct MDPs.

## 2.2 Stochastic Shortest Path on Succinct MDPs

We consider the stochastic shortest path (SSP) problem on succinct MDPs. Below we fix a succinct MDP described by a while-loop in the form (1).

*Reward function.* We consider a reward function $R$ that assigns a reward $R(\mathbf{u}, \ell)$ when the sampling valuation for sampling variables is $\mathbf{u}$ and the non-deterministic choice is $Q_\ell$ (in a loop iteration). We assume that there is a maximal constant $R_{\max} \geq 0$ such that $|R(\mathbf{u}, \ell)| \leq R_{\max}$ for all $\mathbf{u}, \ell$. The rewards need not be nonnegative as our approach is able to handle negative rewards as well. Moreover, our approach can be extended to allow bounded reward functions that depend on program variables, too. See [Chatterjee *et al.*, 2018] for a more detailed discussion.

*Stochastic shortest path.* Given an initial valuation $\mathbf{v}$ for program variables and a policy $\sigma$, the definition of expected total reward/cost until termination is standard. The $\inf$-value (resp., $\sup$-value) of a succinct MDP, given an initial valuation $\mathbf{v}$ for program variables, is the infimum (resp., supremum) expected reward value over all policies that ensure finite expected termination time, which we denote as $\mathit{infval}(\mathbf{v})$ (resp., $\mathit{supval}(\mathbf{v})$).

```
while x ≥ 1 do
□ if (0.4) { x := x + 1 reward=1 }
  else     { x := x − 1 }

□ if (0.3) { x := x + 1 reward=1 }
  else     { x := x − 1 }
```

Figure 1: Gambler's Ruin as a succinct MDP

*Computational problem.* We consider the computational problem of obtaining upper and lower bounds for the $\inf$-value and $\sup$-value for succinct MDPs. Due to the similarity of the problems, we focus on computing lower and upper bounds for the $\sup$-value. The results for $\inf$-value are similar and omitted. For formal details, see [Chatterjee *et al.*, 2018].

## 2.3 Illustrative Examples

**Example 2** (Gambler's Ruin)**.** *We start with a simple and classical example. A gambler has $x_0$ tokens for gambling. In each round he can choose one of the two types of gambling. In type 1, he wins with probability $p_1 < 1/2$ and in type 2 with probability $p_2 < 1/2$. A win leads to a reward of $w$ and an extra token. A loss costs one token. The player gambles as long as he has tokens left. His goal is to maximize overall expected reward. Letting $p_1 = 0.4, p_2 = 0.3$ and $w = 1$, a succinct MDP for this example is shown in Figure 1.*

**Remark 2.** *Note that above we use probabilistic **if** as syntactic sugar, where the assignments in **if**$(p)$ run with probability $p$ and those in the **else** part with probability $1 - p$. Given that the assignments are linear, this can be translated back to a succinct MDP, e.g. the first **if-else** block in Figure 1 is equivalent to:*

$$x := x + r \quad \textbf{reward}{=}0.4$$

*where* r *is a sampling variable with $\mathbb{P}(r = 1) = 0.4$ and $\mathbb{P}(r = -1) = 0.6$.*

**Example 3** (Continuous variant)**.** *We can also consider a continuous variant of the example where the sampling variable $r$ is chosen from some continuous distribution with expected value $-0.2$ (e.g., uniform distribution $[-0.8, 0.4]$).*

# 3 Theoretical Results

In this section we present the main theoretical results which forms the basis of our algorithms of the following section.

*Notation.* Given a succinct MDP in the form (1), we let $X$ be the set of program variables in the program, $|X|$ be the size of $X$, $\mathbb{R}^{|X|}$ be the set of $|X|$-dimensional real-valued vectors, $\mathbf{v} \in \mathbb{R}^{|X|}$ be a valuation for program variables such that the value for the $i$th program variable is the $i$th coordinate of $\mathbf{v}$, $\mathbf{u}$ be a valuation for sampling variables, $N := \{1, \ldots, k\}$

| Notation | Meaning |
|---|---|
| $X$ | the set of program variables |
| $\|X\|$ | the size of $X$ |
| $\mathbb{R}^{\|X\|}$ | the set of $\|X\|$-dimensional real column vectors |
| $\mathbf{v}$ | a valuation for program variables |
| $\mathbf{u}$ | a valuation for sampling variables |
| *infval*$(\mathbf{v})$ | the $\inf$-value for initial valuation $\mathbf{v}$ |
| *supval*$(\mathbf{v})$ | the $\sup$-value for initial valuation $\mathbf{v}$ |
| $N$ | the set of non-deterministic choices |
| $\ell$ | a non-deterministic choice in $N$ |
| $F_\ell$ | the transformation function of $Q_\ell$ mapping valuations before its execution to the resulting valuation after it |
| $R$ | the reward function |
| $R_{\max}$ | an upperbound for the absolute value of the rewards |

Table 1: A Summary of Notation

be the set of non-deterministic choices, $\ell \in N$ be a non-deterministic choice, and $F_\ell$ ($\ell \in N$) be the function such that $F_\ell(\mathbf{v}, \mathbf{u})$ is the valuation for program variables resulting from executing $Q_\ell$ with the valuation $\mathbf{v}$ for program variables and the sampled valuation $\mathbf{u}$ for sampling variables. Table 1 summarizes the notation.

**Upper bounds.** We first introduce the main concept for computing an upper bound for *supval*$(\mathbf{v})$ formally, and then present the informal description.

**Definition 1** (Linear Upper Potential Functions (LUPFs))**.** *A* linear upper potential function (LUPF) *is a function* $h : \mathbb{R}^{|X|} \to \mathbb{R}$ *that satisfies the following conditions:*

(C1) $h$ *is* linear*, i.e., there exist* $\mathbf{a} \in \mathbb{R}^{|X|}$ *and* $b \in \mathbb{R}$ *such that for all* $\mathbf{v} \in \mathbb{R}^{|X|}$*, we have* $h(\mathbf{v}) = \mathbf{a}^{\mathrm{T}} \cdot \mathbf{v} + b$*;*

(C2) *for all* $\mathbf{v}, \mathbf{u}, \ell$ *such that* $\mathbf{v} \models \phi$*,* $F_\ell(\mathbf{v}, \mathbf{u}) \models \neg\phi$ *and* $\ell \in N$*, we have* $K \leq h(F_\ell(\mathbf{v}, \mathbf{u})) \leq K'$ *for some fixed constants* $K$ *and* $K'$*;*

(C3) *for all* $\ell \in N$ *and all valuations* $\mathbf{v}$ *such that* $\mathbf{v} \models \phi$*,*

$$h(\mathbf{v}) \geq \mathbb{E}_{\mathbf{u}}(h(F_\ell(\mathbf{v}, \mathbf{u}))) + \mathbb{E}_{\mathbf{u}}(R(\mathbf{u}, \ell))$$

*where* $\mathbb{E}_{\mathbf{u}}(h(F_\ell(\mathbf{v}, \mathbf{u}))), \mathbb{E}_{\mathbf{u}}(R(\mathbf{u}, \ell))$ *are the expected values over the sampling* $\mathbf{u}$ *when fixing* $\mathbf{v}$ *and* $\ell$*;*

(C4) *for all* $\mathbf{v}$ *such that* $\mathbf{v} \models \phi$*, all sampling valuations* $\mathbf{u}$ *and all* $\ell \in N$*, we have* $|h(\mathbf{v}) - h(F_\ell(\mathbf{v}, \mathbf{u}))| \leq M$ *for some fixed constant* $M \geq 0$*.*

Informally, (C1) specifies the linearity of LUPFs, (C2) specifies that the value of the function at terminating valuations should be bounded, (C3) specifies that the current value of the function at $\mathbf{v}$ is no less than that of the next step at $F_\ell(\mathbf{v}, \mathbf{u})$ plus the cost/reward at the current step, and finally (C4) specifies that the change of value between the current step $\mathbf{v}$ and the next step $F_\ell(\mathbf{v}, \mathbf{u})$ is bounded. Note that $\mathbb{E}_{\mathbf{u}}(h(F_\ell(\mathbf{v}, \mathbf{u})))$ is linear in $\mathbf{v}$ since $h$ and $F_\ell$ are linear. Note that the function $h$ (only) depends on the valuations of the variables before the loop execution, and hence it is only loop-dependent (but not dependent on each assignment).

The following theorem shows that LUPFs indeed serve as upper bounds for *supval*$(\cdot)$.

**Theorem 1.** *If* $h$ *is an LUPF, then* *supval*$(\mathbf{v}) \leq h(\mathbf{v}) - K$ *for all valuations* $\mathbf{v} \in \mathbb{R}^{|X|}$ *such that* $\mathbf{v} \models \phi$*.*

*Proof sketch.* The key ideas of the proof are as follows: Fix any scheduler $\sigma$ that ensures finite expected termination time.

- We first construct a stochastic process based on $h$. Using the condition (C3) which is non-increasing property we establish that the stochastic process obtained is a supermartingale (for definitions of supermartingale see [Williams, 1991, Chapter 10]). The supermartingale in essence preserve the non-increasing property.
- Given the supermartingale, we apply Optional Stopping Theorem (OST) ([Williams, 1991, Chapter 10.10]), and use condition (C4) to establish the required boundedness condition of OST, to arrive at the desired result.

See [Chatterjee *et al.*, 2018] for the full formal proof. □

While conditions (C1) and (C2) are not central to the proof, the condition (C1) ensures linearity, which will be required by our algorithms, and the condition (C2) is the boundedness after termination, that derives the desired upper bounds (i.e., contribution of the term $K$ comes from condition (C2)).

**Lower bounds.** We now consider the lower bounds.

**Definition 2** (Linear Lower Potential Functions (LLPFs))**.** *A* linear lower potential function (LLPF) *is a function* $h : \mathbb{R}^{|X|} \to \mathbb{R}$ *that satisfies (C1),(C2),(C4) and in addition (C3') (instead of (C3)) as follows:*

(C3') *there exists* $\ell \in N$ *such that*

$$h(\mathbf{v}) \leq \mathbb{E}_{\mathbf{u}}(h(F_\ell(\mathbf{v}, \mathbf{u}))) + \mathbb{E}_{\mathbf{u}}(R(\mathbf{u}, \ell))$$

*for all* $\mathbf{v}$ *satisfying* $\mathbf{v} \models \phi$*.*

Similar to Theorem 1, we obtain the following result on lower bounds for *supval*$(\cdot)$.

**Theorem 2.** *If* $h$ *is an LLPF, then* *supval*$(\mathbf{v}) \geq h(\mathbf{v}) - K'$ *for all valuations* $\mathbf{v} \in \mathbb{R}^{|X|}$ *such that* $\mathbf{v} \models \phi$*.*

## 4 Computational Results

By Theorem 1 and Theorem 2, to obtain tight upper and lower bounds for the SSP problem, we need an algorithm to obtain good LUPFs and LLPFs, respectively. We present the results for upper and lower bounds separately.

### 4.1 Computational Approach for Upper Bound

The key steps to obtain an algorithmic approach is as follows: (i) we first establish a linear template with unknown coefficients for a LUPF from (C1); (ii) then we transform logical conditions (C2)–(C4) equivalently into inclusion assertions between polyhedrons; (iii) next we transform inclusion assertions into linear inequalities through Farkas' Lemma; (iv) finally we solve the linear inequalities through linear programming, where the solution for unknown coefficients in the template synthesizes a concrete LUPF that serves as an upper bound for the $\sup$-value. We now recall Farkas' Lemma.

**Theorem 3** (Farkas' Lemma [Farkas, 1894]). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$ *and* $d \in \mathbb{R}$. *Suppose that* $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\} \neq \emptyset$. *Then*

$$\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\} \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^{\mathrm{T}}\mathbf{x} \leq d\}$$

*iff there exists* $\mathbf{y} \in \mathbb{R}^m$ *such that* $\mathbf{y} \geq \mathbf{0}$, $\mathbf{A}^{\mathrm{T}}\mathbf{y} = \mathbf{c}$ *and* $\mathbf{b}^{\mathrm{T}}\mathbf{y} \leq d$.

Intuitively, Farkas' Lemma transforms the inclusion problem of a nonempty polyhedron within a halfspace into a feasibility problem of a system of linear inequalities. As a result, one can decide the inclusion problem in polynomial time through linear programming.

**The Algorithm** UpperBound**.** Consider as input a succinct MDP in the form (1).

1. *Template.* The algorithm sets up a column vector $\mathbf{a}$ of $|X|$ fresh variables and a fresh scalar variable $b$ such that the template for an LUPF $h$ is $h(\mathbf{v}) = \mathbf{a}^{\mathrm{T}} \cdot \mathbf{v} + b$.
2. *Constraints on* $\mathbf{a}$ *and* $b$. We first encode the condition (C2) for the template $h$ as the inclusion assertion

$$\{(\mathbf{v},\mathbf{u}) \mid \mathbf{v} \models \phi, F_\ell(\mathbf{v},\mathbf{u}) \models \neg\phi\} \subseteq \{(\mathbf{v},\mathbf{u}) \mid K \leq \mathbf{a}^{\mathrm{T}} \cdot F_\ell(\mathbf{v},\mathbf{u}) + b \leq K'\}$$

   parameterized with $\mathbf{a}, b, K, K'$ for every $\ell \in N$, where $K, K'$ are fresh unknown constants.
   Then for every $\ell \in N$, the algorithm encodes (C3) as

$$\{\mathbf{v} \mid \mathbf{v} \models \phi\} \subseteq \{\mathbf{v} \mid \mathbf{c}_\ell^{\mathrm{T}} \cdot \mathbf{v} \leq d_\ell\}$$

   where $\mathbf{c}_\ell, d_\ell$ are unique linear combinations of unknown coefficients $\mathbf{a}, b$ satisfying that $\mathbf{c}_\ell^{\mathrm{T}} \cdot \mathbf{v} \leq d_\ell$ is equivalent to $h(\mathbf{v}) \geq \mathbb{E}_{\mathbf{u}}(h(F_\ell(\mathbf{v},\mathbf{u}))) + \mathbb{E}_{\mathbf{u}}(R(\mathbf{u},\ell))$. Finally, the algorithm encodes (C4) as inclusion assertions with a fresh unknown constant $M$ using similar transformations. All the inclusion assertions (with parameters $\mathbf{a}, b, K, K', M$) are grouped *conjunctively* so that these inclusions should all hold.
3. *Applying Farkas' Lemma.* The algorithm applies Farkas' Lemma to all the inclusion assertions generated in the previous step and obtains a system of linear inequalities involving the parameters $\mathbf{a}, b, K, K', M$.
4. *Constraint Solving.* The algorithm calls a linear programming solver on the linear program consisting of the system of linear inequalities generated in the previous step and the minimizing objective function $\mathbf{a}^{\mathrm{T}} \cdot \mathbf{v}_0 + b - K$ where $\mathbf{v}_0$ is an initial valuation for program variables.

*Correctness and running time.* The above algorithm obtains concrete values for $\mathbf{a}, b, K, K', M$ and leads to a concrete LUPF $h$. The correctness that $\mathbf{a}^{\mathrm{T}} \cdot \mathbf{v} + b - K$ is an upper bound for the sup-value follows from Theorem 1. The main optimization solution required by the algorithm is linear programming, and thus our algorithm runs in polynomial time in the size of the input succinct MDP.

**Theorem 4.** *Given a succinct MDP and the SSP problem, the best linear upper bound (wrt an initial valuation) on the sup-value can be computed in polynomial-time in the implicit description of the MDP.*

**Example 4.** *Consider the Gambler's Ruin example (from Section 2.3, Figure 1).*

*Let* $h : \mathbb{R} \to \mathbb{R}$ *be an LUPF for this example, we have:*

$(C1)\ \exists \lambda_1, \lambda_2 \in \mathbb{R}\ \ \forall x \in \mathbb{R}\ \ h(x) = \lambda_1 x + \lambda_2$
$(C2)\ \exists K, K' \in \mathbb{R}\ \ \forall x \in [1,2)\ \ K \leq h(x) \leq K'$
$(C3)\ \forall x \in [1,\infty)\ h(x) \geq 0.4 \cdot (1 + h(x+1)) + 0.6 \cdot h(x-1)$
$(C3)\ \forall x \in [1,\infty)\ h(x) \geq 0.3 \cdot (1 + h(x+1)) + 0.7 \cdot h(x-1)$
$(C4)\ \exists M \in [0,\infty)\, \forall x \in [1,\infty)\ \ |h(x) - h(x-1)| \leq M$
$(C4)\ \ \textbf{\textit{and}}\ \ |h(x) - h(x+1)| \leq M$

*Note that for condition (C2) we need to quantify over* $x \in [1,2)$*, as if* $x$ *is not in this range, then the loop does not terminate in the next iteration. Given condition (C1), the two (C4) conditions are equivalent to* $M \geq \lambda_1$*. Also, the (C2) condition is equivalent to* $K \leq \min\{h(1), h(2)\}$ *and* $K' \geq \max\{h(1), h(2)\}$ *or more precisely* $K \leq \lambda_1 + \lambda_2$*,* $K \leq 2\lambda_1 + \lambda_2$*,* $K' \geq \lambda_1 + \lambda_2$ *and* $K' \geq 2\lambda_1 + \lambda_2$*. By expanding the occurrences of* $h$ *in the first (C3) condition and simplifying, we get* $\forall x \in [1,\infty)\ \ 0 \geq 0.4 - 0.2\lambda_1$ *and we can drop the quantification given that* $x$ *does not appear. Similarly, the second (C3) condition is equivalent to* $0 \geq 0.3 - 0.4\lambda_1$*. In our method, such equivalences are automatically obtained by applying Farkas' Lemma rather than manual inspection of the inequalities. Now that all necessary conditions are replaced by equivalent linear inequalities, we can solve the linear program to find an LUPF. An optimal answer (with minimal* $\lambda_1$*) is the following:* $\lambda_1 = M = 2, \lambda_2 = K = 0, K' = 4$*. Therefore by Theorem 1, we have* $\mathsf{supval}(x_0) \leq 2x_0$ *for all initial valuations* $x_0$ *that satisfy the loop guard.* □

**Remark 3.** *Note that our approach is applicable to succinct MDPs with integer as well as real-valued variables, (i.e., the underlying state-space of the MDP is infinite). Even when we consider integer variables, since* $h$ *gives upper bounds, the reduction is to linear programming, rather than integer linear programming. Note that our approach only depends on expectation of sampling variables, and thus applicable even to continuous sampling variables with same expectation, e.g., our results apply uniformly to Example 2 and Example 3, given that the sampling variable* $r$ *has same expectation.*

## 4.2 Computational Approach for Lower Bound

The algorithm for lower bound is similar to the upper bound, however, there are some subtle and key differences. An important difference is that while in Step 2 of the algorithm for upper bound, there is a *conjunction* of contraints, for the lower bound problem it requires a *disjunction*. This has two important implications: first, we need to consider a generalization of Farkas' lemma and in this case we use Motzkin's Transposition Theorem (which extends Farkas' Lemma with strict inequalities) and second, instead of linear programming we require quadratic programming.

**Theorem 5** (Motzkin's Transposition Theorem [Motzkin, 1936]). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times n}$ *and* $\mathbf{b} \in \mathbb{R}^m, \mathbf{c} \in \mathbb{R}^k$. *Assume that* $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\} \neq \emptyset$. *Then*

$$\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\} \cap \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{B}\mathbf{x} < \mathbf{c}\} = \emptyset$$

*iff there exist* $\mathbf{y} \in \mathbb{R}^m$ *and* $\mathbf{z} \in \mathbb{R}^k$ *such that* $\mathbf{y}, \mathbf{z} \geq \mathbf{0}$, $\mathbf{1}^{\mathrm{T}} \cdot \mathbf{z} > 0$, $\mathbf{A}^{\mathrm{T}}\mathbf{y} + \mathbf{B}^{\mathrm{T}}\mathbf{z} = \mathbf{0}$ *and* $\mathbf{b}^{\mathrm{T}}\mathbf{y} + \mathbf{c}^{\mathrm{T}}\mathbf{z} \leq 0$.

**The Algorithm** LowerBound**.** Our algorithm is similar to UpperBound. For brevity, we only explain the differences.

1. *Template $h$.* Same as in the Algorithm UpperBound.
2. *Constraints on* $\mathbf{a}$ *and* $b$*.* The algorithm first encodes (C2) and (C4) as inclusion assertions in the same way as in UpperBound and transforms them into linear inequalities over $\mathbf{a}, b, K, K', M$ through Farkas' Lemma.
   Then the algorithm transforms (C3') equivalently into the inclusion assertion
   $$\{\mathbf{v} \mid \mathbf{v} \models \phi\} \subseteq \bigcup_{\ell \in N} \{\mathbf{v} \mid \mathbf{c}_\ell^{\mathrm{T}} \cdot \mathbf{v} \leq d_\ell\}$$
   where $\mathbf{c}_\ell, d_\ell$ are determined in the same way as in UpperBound. Furthermore, this inclusion assertion is equivalently written as
   $$\{\mathbf{v} \mid \mathbf{v} \models \phi\} \cap \bigcap_{\ell \in N} \{\mathbf{v} \mid \mathbf{c}_\ell^{\mathrm{T}} \cdot \mathbf{v} > d_\ell\} = \emptyset$$
   and then transformed into a system of quadratic inequalities over $\mathbf{a}$ and $b$ through Motzkin's Transposition Theorem. The system may involve quadratic inequalities since $\mathbf{c}_\ell$ contains the unknown parameters $\mathbf{a}$ and $b$.
3. *Constraint Solving.* The algorithm calls a nonlinear-programming solver on the system of linear and quadratic inequalities generated in the previous step with the maximizing objective function $\mathbf{a}^{\mathrm{T}} \cdot \mathbf{v}_0 + b - K'$ where $\mathbf{v}_0$ is an appropriate initial valuation.

*Correctness and optimization problem.* As for the upper bound, once $\mathbf{a}, b, K, K', M$ are found, we obtain an concrete lower bound $\mathbf{a}^{\mathrm{T}} \cdot \mathbf{v} + b - K'$ for the sup-value from Theorem 2, establishing the correctness of our algorithm. The reduction leads to a quadratic optimization problem of polynomial size wrt the succinct MDP, implying the following result.

**Theorem 6.** *Given a succinct MDP and the SSP problem, the best linear lower bound (wrt an initial valuation) on the sup-value can be computed via a polynomial (in the implicit description of the MDP) reduction to quadratic programming.*

**Example 5.** *Let $h : \mathbb{R} \to \mathbb{R}$ be an LLPF for the Gambler's Ruin example (Figure 1 in Section 2.3). The conditions of the form (C1), (C2) and (C4) are exactly the same as in Example 4. In addition, $h$ must also satisfy the following condition:*

$$(C3') \quad \forall x \in [1, \infty) \quad h(x) \leq 0.4 \cdot (1 + h(x+1)) + 0.6 \cdot h(x-1) \ \textbf{\textit{or}} \ h(x) \leq 0.3 \cdot (1 + h(x+1)) + 0.7 \cdot h(x-1)$$

*Expanding the occurrences of $h$ in the condition above using $h(x) = \lambda_1 x + \lambda_2$, and discarding the quantification, we obtain the following equivalent disjunctive system of inequalities: $0 \leq 0.4 - 0.2\lambda_1$ or $0 \leq 0.3 - 0.4\lambda_1$. This system is obviously equivalent to $\lambda_1 \leq 2$. Note that in general we have disjunction of linear inequalities, which require quadratic programming. As explained previously, such equivalences are automatically obtained by our algorithm using Motzkin's transposition theorem.*

*Adding the equivalent linear forms of conditions (C1), (C2) and (C4) as in Example 4 and considering the resulting linear program with the objective of maximizing $\lambda_1$ leads to the following solution: $\lambda_1 = M = 2, \lambda_2 = -2, K = 0, K' = 2$. Therefore, by Theorem 2, supval$(x_0) \geq 2x_0$ for every initial valuation $x_0$ that satisfies the loop guard. This is the same upper bound we found in Example 4. The bound is tight.* □

# 5 Case Studies and Experiments

## 5.1 Additional Examples

We consider several other classical problems in probabilistic planning that can be described as succinct MDPs.

*Two-dimensional Robot Planning.* Consider a robot that is placed on an initial point $(x_0, y_0)$ of a grid. A player controls the robot and at each step, can order it to move one unit in either direction. However, the robot is not perfect. It follows the order with probability $p < 1$ and ignores it and goes to the left with probability $1 - p$. The process ends when the robot leaves the $x \geq y$ half-plane and the player's objective is to keep the robot in this half-plane. The player gets a reward of 1 each time the robot moves. Note that our method can handle any half-plane and starting point.

*Multi-robot Planning.* Our approach can handle many variables and is only polynomially dependent on the succinct representation of the MDP. To demonstrate this, we consider a scenario similar to the previous case, in which there are now two robots $r_1$ and $r_2$ starting at positions $(x_0, y_0)$ and $(x'_0, y'_0)$. The robot $r_1$ follows the orders with probability $p_1$ and malfunctions and goes right with probability $1 - p_1$. Similarly, $r_2$ follows the commands with probability $p_2$ and goes left with probability $1 - p_2$. The player's goal is to keep $r_1$ to the left of $r_2$, i.e. to keep the robots in the four-dimensional half-space $x \leq x'$. The process ends when the robots leave this half-space and the player gets a reward of 1 per step.

*Mini-roulette.* We model Mini-roulette which is a game on a 13-slot wheel. A player starts with $x_0$ chips. She needs one chip bet and she bets as long as she has chips. If she loses, the chip will not be returned, but a win will not consume the chip and results in a specific amount of (monetary) reward, and possibly even more chips. These types of bets can be placed: (i) *Even money bets:* 6 specific slots are chosen. Then the ball is rolled and the player wins if it lands in one of the 6 slots. So the winning probability is $6/13$. Winning gives a unit reward and one extra chip. (ii) *2-to-1 bets:* 4 slots are chosen and winning gives a reward of 2 and 2 extra chips. (iii,iv,v) *3-to-1, 5-to-1 and 11-to-1 bets:* These correspond to winning probabilities of $3/13, 2/13$ and $1/13$ respectively.

*American Roulette.* The game is like Mini-roulette, except that there are more types of bets and the wheel has 38 slots. The player can now have half-chips. A bet can lead to one of three outcomes: definite loss, partial loss or win. A definite loss costs one chip and a partial loss half a chip. A win pays a reward and more chips. Table 2 summarizes the bets.

## 5.2 Experimental Results

We implemented our approach in Java and experimented on all examples mentioned previously. The results are summa-

| Type | Winning Probability | Partial Loss Probability | Winning Reward | Winning Tokens |
|---|---|---|---|---|
| 35-to-1 | 1/38 | 0 | 35 | 35 |
| 17-to-1 | 1/19 | 0 | 17 | 17 |
| 11-to-1 | 3/38 | 0 | 11 | 11 |
| 8-to-1 | 2/19 | 0 | 8 | 8 |
| 6-to-1 | 5/38 | 0 | 6 | 6 |
| 5-to-1 | 3/19 | 0 | 5 | 5 |
| 2-to-1 | 6/19 | 1/19 | 2 | 2 |
| Even | 9/19 | 1/19 | 1 | 1 |

Table 2: Types of bets available in American Roulette.

| MDP | Parameters | Upper bound | Lower bound | Time |
|---|---|---|---|---|
| Gambler's Ruin | $p_1 = 0.4$ $p_2 = 0.3$ | $2x$ | $2x$ | 153 ms |
| 2D Robot Planning | $p = 0.4$ | $5x - 5y$ | $5x - 5y$ | 251 ms |
| Multi-robot Planning | $p_1 = 0.4$ $p_2 = 0.4$ | $2.5x - 2.5y + 5$ | $2.5x - 2.5y$ | 758 ms |
| Mini-roulette | — | $11x$ | $11x$ | 320 ms |
| American Roulette | — | $24x$ | $24x$ | 425 ms |

Table 3: Experimental Results

rized in Table 3, where "Parameters" shows concrete parameters for our examples, "Upper bound" (resp. "Lower bound") presents the LUPFs (resp. LLPFs) obtained through our approach, and finally "Time" shows the running time in milliseconds. The reported upper bounds on sup-values are the results of $h(\mathbf{v}) - K$ as in Theorem 1. Similarly, lower bounds on sup-values are obtained from $h(\mathbf{v}) - K'$ as in Theorem 2. Finally, our approach is not sensitive to parameters as varying them will only change coefficients of our LUPFs/LLPFs.

*Runtime.* Our approach is extremely efficient and handles all these MDPs, even when the succinct representation is large, in less than a second. The results were obtained on an Intel Core i5-2520M machine, running Ubuntu. We used lp-solve [Berkelaar *et al.*, 2004], JavaILP [Lukasiewycz, 2008] and JOptimizer [Tivellato, 2017] for optimization tasks.

*Significance of our results.* First, observe that in most experimental results the upper and lower bounds are tight. Thus our approach provides precise bounds on the SSP problem for several classical examples. Second, our results apply to infinite-state MDPs: in all the above examples, we consider infinite-state MDPs, where algorithmic approaches for factored MDPs do not work. Finally, in the above examples, instead of infinite-state MDPs if we consider large finite-state MDP variants (e.g., bounding $x$ in Gambler's Ruin with a large domain), then as the MDP becomes larger, the SSP value of the finite-state MDP approaches the infinite-state value. Hence, our tight bounds on the infinite-state SSP value provide efficient approximation for large finite-state MDPs.

## 6 Related Works

*MDPs.* MDPs are widely studied in the AI literature [Sigaud and Buffet, 2010; Dean *et al.*, 1997; Singh *et al.*, 1994; Williams and Young, 2007; Poupart *et al.*, 2015; Gilbert *et al.*, 2017; Topin *et al.*, 2015; Perrault and Boutilier, 2017; Boutilier and Lu, 2016; Ferns *et al.*, 2004]; and factored MDPs are considered as an effective approach [Guestrin *et al.*, 2003]. We introduced succinct MDPs and efficient algorithms which are applicable to several problems in AI.

*PPDDL and RDDL.* There are a variety of languages for specifying MDPs and especially factored MDPs. Two of the most commonly used are PPDDL [Younes and Littman, 2004] and RDDL [Sanner, 2010]. These are general languages capturing all factored MDPs. Instead, we consider succinct MDPs with linear guards and assignments and no nested while-loops. Hence, our language is simpler than them.

*Programming language results.* Besides the AI community, research in programming languages also considers probabilistic programs and algorithmic approaches [Chakarov and Sankaranarayanan, 2013; Chatterjee *et al.*, 2016]; but the main focus is on termination, whereas we consider the SSP and compute precise bounds. While both approaches use the theory of martingales, the key differences are as follows:

- *Problem difference:* [Chatterjee *et al.*, 2016] considers the number of steps to termination. There is no notion of reward or stochastic shortest path. In contrast, we consider rewards and SSP. In particular, we have negative rewards that cannot be modeled by the notion of steps.
- *Result difference:* [Chatterjee *et al.*, 2016] considers the qualitative question of whether expected termination time is finite or not, and then applies Azuma's inequality to martingales for concentration bounds. In contrast, we present upper and lower bounds on expected SSP. Thus our results provide quantitative (rather than qualitative) bounds for expected SSP that significantly generalize expected termination time. However, our results are applicable to a more restricted class of programs.
- *Proof-technique difference:* [Chatterjee *et al.*, 2016] considers qualitative expected termination time characterization, and the main mathematical tool is martingale convergence that does not handle negative rewards. In contrast, we present quantitative bounds for SSP and our mathematical tool is Optional Stopping Theorem.

*Comparison with [Hansen and Abdoulahi, 2015].* This work provides convergence tests for heuristic search value-iteration algorithms. While both approaches provide bounds for SSPs, the main differences are as follows: (i) our approach can handle negative costs whereas [Hansen and Abdoulahi, 2015] can handle only positive costs; (ii) our results are on the implicit representation of MDPs, while [Hansen and Abdoulahi, 2015] evaluates parts of the explicit MDP; and (iii) our approach presents polynomial reductions to optimization problems and is not dependent on value-iteration.

## 7 Conclusion

We consider succinct MDPs and present algorithms for the SSP problem on them. A direction for future work is to consider other algorithmic approaches for succinct MDPs. Also, we consider linear templates and extending our approach to more general templates is another interesting direction.

## Acknowledgements

The research was partially supported by Vienna Science and Technology Fund (WWTF) Project ICT15-003, Austrian Science Fund (FWF) NFN Grant No S11407-N23 (RiSE/SHiNE), and ERC Starting grant (279307: Graph Games).

## References

[Berkelaar *et al.*, 2004] Michel Berkelaar, Kjell Eikland, Peter Notebaert, et al. LPsolve: Open source (mixed-integer) linear programming system. Technical report, 2004.

[Bertsekas, 2005] Dimitri Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 3rd edition, 2005.

[Boutilier and Lu, 2016] Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained Markov decision processes. In *UAI*, pages 52–61, 2016.

[Chakarov and Sankaranarayanan, 2013] Aleksandar Chakarov and Sriram Sankaranarayanan. Probabilistic program analysis with martingales. In *CAV*, pages 511–526, 2013.

[Chatterjee *et al.*, 2016] Krishnendu Chatterjee, Hongfei Fu, Petr Novotnỳ, and Rouzbeh Hasheminezhad. Algorithmic analysis of qualitative and quantitative termination problems for affine probabilistic programs. In *POPL*, pages 327–342, 2016.

[Chatterjee *et al.*, 2018] Krishnendu Chatterjee, Hongfei Fu, Amir Goharshady, and Nastaran Okati. Computational approaches for stochastic shortest path on succinct MDPs. *arXiv preprint arXiv:1804.08984*, 2018.

[Dean *et al.*, 1997] Thomas Dean, Robert Givan, and Sonia Leach. Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In *UAI*, pages 124–131, 1997.

[Delgado *et al.*, 2011] Karina Valdivia Delgado, Scott Sanner, and Leliane Nunes De Barros. Efficient solutions to factored MDPs with imprecise transition probabilities. *Artif. Intell.*, 175(9-10):1498–1527, 2011.

[Farkas, 1894] Julius Farkas. A Fourier-Féle mechanikai elv alkalmazásai (Hungarian). *Mathematikaiés Természettudományi Értesitö*, 12:457–472, 1894.

[Ferns *et al.*, 2004] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, pages 162–169, 2004.

[Filar and Vrieze, 1997] Jerzy Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer, 1997.

[Gilbert *et al.*, 2017] Hugo Gilbert, Paul Weng, and Yan Xu. Optimizing quantiles in preference-based Markov decision processes. In *AAAI*, pages 3569–3575, 2017.

[Guestrin *et al.*, 2003] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *J. Artif. Intell. Res.*, 19:399–468, 2003.

[Hansen and Abdoulahi, 2015] Eric A Hansen and Ibrahim Abdoulahi. Efficient bounds in heuristic search algorithms for stochastic shortest path problems. In *AAAI*, pages 3283–3290, 2015.

[Howard, 1960] Ronald A Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[Lukasiewycz, 2008] Martin Lukasiewycz. JavaILP - java interface to ILP solvers, javailp.sourceforge.net, 2008.

[Motzkin, 1936] Theodore Samuel Motzkin. *Beiträge zur Theorie der linearen Ungleichungen (German)*. PhD thesis, Basel, Jerusalem, 1936.

[Perrault and Boutilier, 2017] Andrew Perrault and Craig Boutilier. Multiple-profile prediction-of-use games. In *AAMAS*, pages 275–295, 2017.

[Poupart *et al.*, 2015] Pascal Poupart, Aarti Malhotra, Pei Pei, Kee-Eung Kim, Bongseok Goh, and Michael Bowling. Approximate linear programming for constrained partially observable Markov decision processes. In *AAAI*, pages 3342–3348, 2015.

[Puterman, 1994] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[Sanner, 2010] Scott Sanner. Relational dynamic influence diagram language (RDDL): Language description. Technical report, 2010.

[Sigaud and Buffet, 2010] Olivier Sigaud and Olivier Buffet. *Markov Decision Processes in Artificial Intelligence*. Wiley, 2010.

[Singh *et al.*, 1994] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Learning without state-estimation in partially observable Markovian decision processes. In *Machine Learning Proceedings*, pages 284 – 292. Morgan Kaufmann, 1994.

[Tivellato, 2017] Alberto Tivellato. JOptimizer - Java convex optimizer, www.joptimizer.com, 2017.

[Topin *et al.*, 2015] Nicholay Topin, Nicholas Haltmeyer, Shawn Squire, John Winder, Marie desJardins, and James MacGlashan. Portable option discovery for automated learning transfer in object-oriented Markov decision processes. In *IJCAI*, pages 3856–3864, 2015.

[Williams and Young, 2007] Jason D Williams and Steve Young. Partially observable Markov decision processes for spoken dialog systems. *Comput. Speech Lang.*, 21(2):393–422, 2007.

[Williams, 1991] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[Younes and Littman, 2004] Håkan Younes and Michael L Littman. PPDDL1. 0: The language for the probabilistic part of IPC-4. In *IPC*, pages 70–74, 2004.