In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann: San Francisco, CA 61-68, August 2002.

TECHNICAL REPORT
R-304
August 2002

Qualitative MDPs and POMDPs: An Order-of-Magnitude Approximation

Blai Bonet and **Judea Pearl**
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90024
{bonet,judea}@cs.ucla.edu

Abstract

We develop a qualitative theory of Markov Decision Processes (MDPs) and Partially Observable MDPs that can be used to model sequential decision making tasks when only qualitative information is available. Our approach is based upon an order-of-magnitude approximation of both probabilities and utilities, similar to ϵ-semantics. The result is a qualitative theory that has close ties with the standard theory arithmetics and is amenable to the most recent general planning techniques.

1 Introduction

The general task of sequential decision making under uncertainty and partial information is of central importance in AI since it embraces a broad range of common problems found in planning, robotics, game solving, etc. Currently, the most general and clear formulation of the task is achieved through the theory of Markov Decision Processes (MDPs) and Partially Observable MDPs (POMDPs) [8, 16, 4]. These models do not only provide a sound, concise and general framework for modeling complex problems but also algorithms for solving them, the most important being the Value Iteration and Policy Iteration algorithms.

The standard formulation for an MDP (or POMDP) model is made of two types of ingredients:

(a) qualitative information that define the structure of the problem, e.g. the set of world configurations (state space), the set of available decisions (also known as controls or actions), the feedback that can be received by the agent, etc. and

(b) quantitative information (also known as parametrization of the structure) that, together with the qualitative information, defines the model. Typical examples are the transition probabilities of going from one state to another after the application of a control, the costs incurred in such application, etc.

In general, the (optimal) solution to an MDP (POMDP) depends in both types of information, so the standard algorithms need such information. Quite often, however, we have precise knowledge of the qualitative information but only "rough" estimates of the quantitative parameters. In such cases, the standard algorithms cannot be applied unless the missing information is "completed". A process that is arbitrary and, more important, sometimes unnecessary for obtaining reasonable solutions. Thus, we would like to have a well-founded framework in which partially specified problems can be solved.[1]

In this paper, we develop a *qualitative theory* for MDPs and POMDPs in which such underspecified problems can be modeled and solved. As it will be seen, the resulting theory can be thought as a generalization of the standard MDP theory in the sense that as the quantitative information becomes more precise, the qualitative processes become "closer" to the standard processes.

In the computational side, we will show how the Value Iteration algorithm can be used as a general algorithm for solving the qualitative models. Yet for some subclasses of problems, other recent algorithms might work better, specially those based on heuristic search, symbolic model-checking and SAT.

The paper is organized as follows. In the next Section, we review the formal definitions and most important results of the standard theory of MDPs and POMDPs. In Sect. 3, we present the formal foundations upon which the qualitative theory of MDPs and POMDPs is built. The qualitative theory of MDPs and POMDPs is given in Sections 4 and 5 respectively. Then, in Sect. 6 we discuss and give pointers for the computational issues that arise from special subclasses of problems. We

[1] Other approaches for solving unspecified MDPs are those based on Reinforcement Learning, yet they basically estimate the quantitative information from experimentation; see [26].

finish the paper with a brief discussion that includes related work and a summary. Due to space limitations, we only provide proofs for the most novel results.

2 Standard MDPs and POMDPs

In this section and the rest of the paper we use a notation similar to the one in [3]; the reader is referred there for an excellent introduction to MDPs.

The MDP model assumes the existence of a physical system that evolves in discrete time and that is controlled by an agent. The system dynamics is governed by probabilistic transition functions that maps states and controls to states. At every time, the agent incurs in a cost that depends in the current state of the system and the applied control. Thus, the task is to find a control strategy (also known as a policy) that minimize the expected total cost over the infinite horizon time setting. Formally, an MDP is characterized by

(M1) A finite state space $S = \{1, \ldots, n\}$,

(M2) a finite set of controls $U(i)$ for each state $i \in S$,

(M3) transition probabilities $p_{i,u}(j)$ for all $u \in U(i)$ that are equal to the probability of the next state being j after applying control u in state i, and

(M4) a cost $g(i, u)$ associated to $u \in U(i)$ and $i \in S$.

A strategy or policy π is an infinite sequence $(\mu_0, \mu_1, \ldots)$ of functions where μ_k maps states to controls so that the agent applies the control $\mu_k(i)$ in state $x_k = i$ at time k, the only restriction being that $\mu_k(i) \in U(i)$ for all $i \in S$. If $\pi = (\mu, \mu, \ldots)$, the policy is called *stationary* (i.e. the control does not depend in time) and is simply denoted by μ. The cost associated to π when the system starts at state x_0 is:

$$J_\pi(x_0) \doteq \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k))\right\} \qquad (1)$$

where the expectation is taken with respect to the probability distribution induced by the transition probabilities, and where the number $\alpha \in [0, 1]$, called the *discount factor*, is used to discount future costs at a geometric rate.

The MDP *problem* is to find an *optimal policy* π^* satisfying $J^*(i) \doteq J_{\pi^*}(i) \leq J_\pi(i)$ $(i = 1, \ldots, n)$, for every other policy π. Although there could be none or more than one optimal policy, the optimal cost vector J^* is always unique. The case $\alpha < 1$ is of outmost importance since it guarantees that the optimal policy always exists and, more important, that there exists a stationary policy that is optimal. In such case, J^* is the unique solution to the *Bellman* equation:

$$J^*(i) = \min_{u\in U(i)} g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J^*(j). \qquad (2)$$

Also, if J^* is a solution for (2) then the *greedy* stationary policy μ^* with respect to J^*:

$$\mu^*(i) \doteq \operatorname*{argmin}_{u\in U(i)} \left\{ g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J^*(j) \right\} \qquad (3)$$

is an optimal and stationary policy for the MDP. Therefore, solving the MDP is equivalent to solving (2). Such equation can be solved using the DP *operator*:

$$(TJ)(i) \doteq \min_{u\in U(i)} g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J(j) \qquad (4)$$

that map $\mathbb{R}^n$ into $\mathbb{R}^n$. When $\alpha < 1$ it is not hard to show that the DP operator is a contraction mapping with fixed point J^* that satisfy:[2]

$$J^* = TJ^* = \lim_{k\to\infty} T^k J_0 \qquad (5)$$

where J_0 is the zero n-dimensional vector. The Value Iteration algorithm computes J^* iteratively by using (2) as an update rule. Starting from any vector J, Value Iteration computes a succession of vectors $\langle J_k \rangle_{k\geq 0}$ defined by $J_0 \doteq J$ and $J_{k+1} \doteq TJ_k$. The algorithm stops when $J_{k+1} = J_k$, or when the residual $\max_{i\in S} |J_{k+1}(i) - J_k(i)|$ is sufficiently small. In the latter case, the suboptimality of the resulting policy is bounded by a constant multiplied by the residual.

2.1 Partially Observable MDPs

A Partially Observable Markov Decision Process (POMDP) is an MDP in which the agent does not know the state of the system. This is an important departure from the MDP model since even if the agent knows the optimal strategy for the underlying MDP, it cannot apply it. Thus, the agent needs to estimate the state of the system and then act accordingly. Such estimates are known as the *belief states* of the agent and are updated continuously as the system evolves. The POMDP framework alse extends the MDP framework by allowing controls to return information about the system. For example, a `blood-test` might return blood type and `reading-radar` might return the distance to objects. See [6] for definitions from the AI perspective. Formally, a POMDP is characterized by:

(P1) A finite state space $S = \{1, \ldots, n\}$,

(P2) a finite set of controls $U(i)$ for each state $i \in S$,

(P3) transition probabilities $p_{i,u}(j)$ for all $u \in U(i)$ equal to the probability of the next state being j after applying u in i,

(P4) a finite set of observations $O(i, u) \subseteq O$ that may result after applying $u \in U(i)$ in $i \in S$,

[2] A contraction mapping $T : S \to S$ over a Banach space S with norm $\|\cdot\|$ is a bounded operator such that $\|TJ - TJ'\| < \alpha\|J - J'\|$ with $\alpha < 1$. In this case, it is known that T is continuous, that has a unique fixed point J^*, and that $T^n J \to J^*$ as $n \to \infty$ for any $J \in S$. See [18].

(P5) observation probabilities $p_{i,u}(o)$ for all $u \in U(i)$ and $o \in O(i,u)$ equal to the probability of receiving o in i after applying u, and

(P6) a cost $g(i,u)$ associated to $u \in U(i)$ and $i \in S$.

It had been shown that finding an optimal strategy to this problem is equivalent to solving an *infinite* MDP problem in *belief space*, the so-called belief-MDP, whose elements are:

(B1) A belief space B of prob. distributions over S,

(B2) a set of controls $U(x) = \{u : \forall i[x(i) > 0 \Rightarrow u \in U(i)]\}$ for each belief state $x \in B$, and

(B3) a cost $g(x,u) = \sum_{i=1}^{n} g(i,u)\,x(i)$ for each $u \in U(x)$ and $x \in B$.

The transition probabilities of the belief-MDP are determined by the abilities of the agent. It is known that a full capable and rational agent should perform *Bayesian* updating in order to behave optimally. In that case, the transition probabilities are

$$x \rightsquigarrow x_u^o \quad \text{with probability} \quad p_{x,u}(o)$$

where $u \in U(x)$, $o \in O(x,u)$ the set of possible observations after applying control u in belief state x, $p_{x,u}(o)$ is the probability of receiving observation o after applying u in x, and x_u^o is the Bayesian update of x after u and o; i.e.

$$x_u^o(i) \doteq \frac{x_u(i)\,p_{i,u}(o)}{p_{x,u}(o)}, \tag{6}$$

$$x_u(i) \doteq \sum_{j=1}^{n} x(j)\,p_{j,u}(i), \tag{7}$$

$$p_{x,u}(o) \doteq \sum_{i=1}^{n} x_u(i)\,p_{i,u}(o), \tag{8}$$

$$O(x,u) \doteq \{o : p_{x,u}(o) > 0\}. \tag{9}$$

And the corresponding DP operator is:

$$(TJ)(x) \doteq \min_{u \in U(x)} \; g(x,u) + \alpha \sum_{o \in O(x,u)} p_{x,u}(o) J(x_u^o).$$

As in the MDP case, when $\alpha < 1$, the DP operator is a contraction mapping so it has a unique fixed point that is the solution to the Bellman equations. This fact guarantees the existence of an stationary strategy that is optimal but, unfortunately, the Value and Policy Iteration algorithms are no longer feasible (but see [15] for a survey of POMDPs algorithms).

3 Foundations

Our approach towards a qualitative theory for MDPs and POMDPs will be based on the qualitative decision theory proposed by Wilson few years ago [27]. Wilson's theory, built upon ideas of Pearl and Goldszmidt [22, 21, 14], defines a set of abstract quantities called *Extended Reals*, denoted by $\mathcal{Q}$, that are used to represent qualitative probabilities and utilities. Each Extended Real is a *rational function* p/q where p and q are polynomials in ε with coefficients in the rationals. Plainly, ε is thought as a very small but unknown quantity so that the Extended Reals can be used to represent "information up to ε precision." For example, quantities like $1-\varepsilon$ and ε might be used for the qualitative probabilities "likely" and "unlikely" respectively, and ε^{-1} for a high utility. These quantities are then combined using standard arithmetic operations between polynomials for computing expected qualitative utilities. The utilities are then compared among each other by means of a linear order $\succeq$ in $\mathcal{Q}$ that is defined in [27].[3]

In order to use Wilson's theory to define a qualitative version of MDPs we need to be sure that equations like (1) and (5) are well-defined. That is, we need to define some notion of convergence and then give conditions that guarantee the existence of such limits.

Let us begin with some intuition about the problem we face. Consider the sequence $\langle s_n \rangle_{n\geq 0}$ defined by $s_n = \sum_{k=0}^{n} \varepsilon^k$. Since the difference between two consecutive elements $s_{n+1} - s_n = \varepsilon^{n+1}$, we would expect that such sequence "converges" to $\sum_{k\geq 0} \varepsilon^k$. This notion of convergence is what we are after. As the reader may observe the set of candidate limits must include some collection of series in ε.

3.1 A Complete Extension of $\mathcal{Q}$

Consider the set $\mathcal{S}_\rho$ of *two-sided infinite formal series* in ε with real coefficients $s = \sum_k a_k \varepsilon^k$ such that

$$\|s\|_\rho \doteq \sum_k |a_k| \rho^{-k} \; < \; \infty,$$

with $\rho > 0$ a real number, where the summation is over all integers. We say that a sequence $\langle s_n \rangle_{n\geq 0}$ from $\mathcal{S}_\rho$ *converges* to $s \in \mathcal{S}_\rho$ iff $\|s - s_n\|_\rho \to 0$ as $n \to \infty$, and that the sequence is *convergent* iff $\|s_n - s_m\|_\rho \to 0$ as $n, m \to \infty$. Note that there is no reason for a convergent sequence to converge to something in $\mathcal{S}_\rho$. Fortunately, this is not the case. This fact is known as that $\langle \mathcal{S}_\rho, \|\cdot\|_\rho \rangle$ is a *complete* normed space (i.e. Banach).[4] We define the following arithmetic operations of sum, multiplication by an scalar and multiplication between

[3] More precisely, Wilson's paper defines $f(\varepsilon) \succeq g(\varepsilon)$ if and only if $f(\varepsilon) - g(\varepsilon) \geq 0$ for all $\varepsilon \in (0, \xi)$ where $\xi > 0$ depends in f and g.

[4] Indeed, $\langle \mathcal{S}_\rho, \|\cdot\|_\rho \rangle$ is the $L_1(\mu_\rho)$ space with respect to the measure μ_ρ over Z defined by $\mu_\rho\{k\} = \rho^{-k}$. The Riesz-Fischer Theorem in Analysis asserts that this space is complete (i.e. it is a Banach Space) [23, Ch.11.§7].

series:

$$(s+t)(k) \doteq s(k)+t(k),$$
$$(\alpha s)(k) \doteq \alpha s(k),$$
$$(s\cdot t)(k) \doteq \sum_{i,j}[\![i+j=k]\!]s(i)t(j) = \sum_i s(i)\,t(k-i)$$

where $s(k)$ is the functional notation for the kth term of s, and $[\![\varphi]\!]$ is 1 (resp. 0) if φ is true (resp.false).

Two important elements in $\mathcal{S}_\rho$ are $\mathbf{0}$ and $\mathbf{1}$ defined as $\mathbf{0}(k) \equiv 0$ and $\mathbf{1}(k) = 1$ (resp. 0) if $k = 0$ (resp. $k \neq 0$). They are respectively the identity elements for the sum and product of series. The *order-of-magnitude* of $s \in \mathcal{S}_\rho$ is defined by $s^\circ \doteq \inf\{k \in \mathbb{Z} : s(k) \neq 0\}$ and $\mathbf{0}^\circ \doteq \infty$. For the subset $\underline{\mathcal{S}}_\rho$ of series such that $s^\circ > -\infty$ we have

Theorem 1 *$\underline{\mathcal{S}}_\rho$ is a field; i.e. $(\underline{\mathcal{S}}_\rho, \mathbf{0}, +)$ is a commutative group, $(\underline{\mathcal{S}}_\rho \setminus \{\mathbf{0}\}, \mathbf{1}, \cdot)$ is a commutative group, and the product is distributive with respect to the sum.*

Example 1: Let $s = \sum_{k\geq 0} 2^{-k}\varepsilon^k$. Then, its multiplicative inverse is $s^{-1} = 1 - \frac{1}{2}\varepsilon$. To check this, for $k > 0$,

$$\begin{aligned}
(s\cdot s^{-1})(-k) &= 0,\\
(s\cdot s^{-1})(0) &= s(0)s^{-1}(0) = 1,\\
(s\cdot s^{-1})(k) &= \sum_i s(i)s^{-1}(k-i)\\
&= s(k-1)s^{-1}(1) + s(k)s^{-1}(0)\\
&= -2^{-k+1}2^{-1} + 2^{-k} = 0.
\end{aligned}$$

Hence $s \cdot s^{-1} = \mathbf{1}$. □

It is not hard to show that the set of Extended Reals $\mathcal{Q}$ is *dense* in $\mathcal{S}_\rho$. That is, that for any $s \in \mathcal{S}_\rho$ there exists a sequence $\langle s_n\rangle_{n\geq 0}$ from $\mathcal{Q}$ such that $s_n \to s$.

3.2 Linear Order in $\underline{\mathcal{S}}_\rho$

We define a linear order in $\underline{\mathcal{S}}_\rho$ from the linear order in $\mathcal{Q}$ using the fact that $\mathcal{Q}$ is dense. The construction is done in the standard way by defining the set $\mathcal{P}$ of positive elements in $\underline{\mathcal{S}}_\rho$.

Let us denote with $\succeq$ the order in $\mathcal{Q}$ and let $s \in \underline{\mathcal{S}}_\rho$ be not equal to $\mathbf{0}$. Since $\mathcal{Q}$ is dense, there exists a sequence $\langle s_n\rangle_{n\geq 0}$ from $\mathcal{Q}$ that converges to s. Moreover, we can choose the series such that ${s_n}^\circ \leq s^\circ$ for all n. We say that $s \in \mathcal{P}$ if and only if there exists an integer N such that $s_n \succ \mathbf{0}$ for all $n > N$.The following result shows that $\mathcal{P}$ is well-defined and satisfies the desired conditions.

Theorem 2 *Let $s \in \underline{\mathcal{S}}_\rho$ be different from $\mathbf{0}$. Then*

(a) s is in (or not in) $\mathcal{P}$ independently of the chosen series $\langle s_n\rangle$, i.e. from any two such series $\langle s_n\rangle$ and $\langle s'_n\rangle$ the conclusion is the same,

(b) either $s \in \mathcal{P}$ or $-s \in \mathcal{P}$,

(c) $s \in \mathcal{P}$ if and only if $s(s^\circ) > 0$.

For $s,t \in \mathcal{S}_\rho$ we say that $s > t$ if and only if $s - t$ is in $\mathcal{P}$, the other relations $<, \geq, \ldots$ are defined in the usual way. Unfortunately the field $\underline{\mathcal{S}}_\rho$ also lacks, as $\mathcal{Q}$, the *least upper bound* property of the reals, that is that every bounded set has a least upper bound (similarly for greatest lower bounds).

3.3 Normed Vector Spaces

An n-dimensional normed vector space with elements in $\mathcal{S}_\rho$ can be defined using the norm $\|J\| \doteq \sup_{i=1,\ldots,n}\|J(i)\|_\rho$. This space, denoted by $\mathcal{S}_\rho^n$, is a normed real vector space. Since $\mathcal{S}_\rho$ is complete $\mathcal{S}_\rho^n$ is also complete (i.e. Banach). A map $T : \mathcal{S}_\rho^n \to \mathcal{S}_\rho^n$ is a contraction mapping with coefficient $\alpha \in [0,1)$ if $\|TJ - TH\| \leq \alpha\|J - H\|$ for all $J, H \in \mathcal{S}_\rho^n$. If so, T has a unique fix point $J^* \in \mathcal{S}_\rho^n$.

We also will deal with more general vector spaces whose elements can be though as mappings $\mathcal{S}_\rho^X$ where X is a (finite or infinite) set. The corresponding norm is $\|J\| = \sup_{x\in X}\|J(x)\|_\rho$. Thus, if $|X| = n$, $\mathcal{S}_\rho^X$ is the n-dimensional space $\mathcal{S}_\rho^n$ and if $|X| = \infty$, then $\mathcal{S}_\rho^X$ is infinite dimensional.

3.4 $\mathcal{S}_\rho$-Probability Spaces

An $\mathcal{S}_\rho$-probability space is a triplet $(\Omega, \mathcal{F}, P)$ where Ω is a *finite* set of outcomes, $\mathcal{F}$ is the set of all subsets of Ω, and P is a $\mathcal{S}_\rho$-valued function on $\mathcal{F}$ such that:

(a) $P(A) \geq \mathbf{0}$ for all $A \subseteq \Omega$,

(b) $P(A \cup B) = P(A) + P(B)$ for all disjoint $A, B \subseteq \Omega$,

(c) $P(\Omega) = \mathbf{1}$.

In this case we say that P is a $\mathcal{S}_\rho$-probability over Ω, or that P is a *qualitative probability* over Ω. A random variable X on $(\Omega, \mathcal{F}, P)$ is a mapping $\Omega \to \mathcal{S}_\rho$. Its expected value is defined as $EX \doteq \sum_{\omega\in\Omega} X(\omega)\,P(\{\omega\})$.

3.4.1 Kappa Rankings

A kappa ranking is a function κ that maps subsets in $\mathcal{F}$ into the non-negative integers (including ∞) such that: $\kappa(\emptyset) = \infty$, $\kappa(\Omega) = 0$ and $\kappa(A \cup B) = \min\{\kappa(A), \kappa(B)\}$ for disjoint $A, B \subseteq \Omega$. This function should be thought as a ranking of the set of worlds into degrees of *disbelief* so that values of 0,1,2, ... refer to situations deemed as probable, unprobable, very unprobable, ... Such values have a close connection with qualitative probabilities since if P is a qualitative probability, then P° is a kappa ranking. Kappa rankings had been used to model order-of-magnitude approaches to decision making [25, 14, 21] and are closely related to other qualitative approaches like [9]. It is easy to see that

$\kappa : \Omega \to \mathbb{Z}^+ \cup \{\infty\}$ is a kappa ranking iff there exists $\omega \in \Omega$ such that $\kappa(\omega) = 0$.

As just said, any qualitative probability induces a kappa ranking over worlds. The other direction, going from kappa rankings to qualitative probabilities, is not uniquely determined. However, we define one such mapping, called *projection*, that will play a major role when defining the qualitative processes.

3.4.2 Projections

A projection ζ_κ of a kappa ranking κ over (finite) Ω is a qualitative probability over Ω that is *consistent* with κ, i.e. $\zeta_\kappa(\omega)^\circ = \kappa(\omega)$ for all $\omega \in \Omega$. To define ζ_κ, let n_m be the number of elements from Ω such that κ assign m to them, i.e. $|\{\omega \in \Omega : \kappa(\omega) = m\}| = n_m$. Then, ζ_κ is defined as $\zeta_\kappa(\omega) = \mathbf{0}$ if $\kappa(\omega) = \infty$ and

$$\zeta_\kappa(\omega) \doteq \frac{N_{\kappa(\omega)}}{n_{\kappa(\omega)}} \left(\varepsilon^{\kappa(\omega)} - \sum_{j>\kappa(\omega)} [\![n_j \neq 0]\!]\, \varepsilon^j \right) \qquad (10)$$

otherwise, where the integers N_k are defined as

$$N_0 \doteq 1, \quad N_k \doteq \sum_{j=0}^{k-1} [\![n_j \neq 0]\!]\, N_j .$$

Later, in Sect. 7, we will discuss the "arbitrariness" of this definition.

Example 2: Let $\Omega = \{a, b, c, d, e\}$ and κ be such that $\kappa(a) = \kappa(b) = 0$, $\kappa(c) = \kappa(d) = 1$ and $\kappa(e) = 5$. Then,

$$\begin{aligned} \zeta_\kappa(a) &= \zeta_\kappa(b) = 2^{-1}\left(1 - \varepsilon - \varepsilon^5\right), \\ \zeta_\kappa(c) &= \zeta_\kappa(d) = 2^{-1}\left(\varepsilon - \varepsilon^5\right), \\ \zeta_\kappa(e) &= 2\varepsilon^5. \end{aligned}$$

Clearly, $\zeta_\kappa(a) + \zeta_\kappa(b) + \zeta_\kappa(c) + \zeta_\kappa(d) + \zeta_\kappa(e) = \mathbf{1}$ so ζ_κ is a qualitative probability consistent with κ. □

Theorem 3 *Let κ be a kappa ranking over Ω. Then,*

(a) ζ_κ is a qualitative proability over Ω,

(b) $\zeta_\kappa(\omega)^\circ = \kappa(\omega)$ for all $\omega \in \Omega$, and

(c)
$$\sup_\kappa \sum_{\omega\in\Omega} \|\zeta_\kappa(\omega)\|_\rho \to 1 \quad \text{as} \quad \rho \to \infty,$$

where the sup is over all kappa rankings over Ω.

Proof: Part (b) follows readily from definitions. For (a), we only need to show that $\sum_\omega \zeta_\kappa(\omega) = \mathbf{1}$:

$$\begin{aligned} &\sum_{\omega\in\Omega} \zeta_\kappa(\omega) \\ &= \sum_{\omega\in\Omega} [\![\kappa(\omega) \neq \infty]\!] \frac{N_{\kappa(\omega)}}{n_{\kappa(\omega)}} \left(\varepsilon^{\kappa(\omega)} - \sum_{j>\kappa(\omega)} [\![n_j \neq 0]\!]\, \varepsilon^j \right) \\ &= \sum_{k\geq 0} [\![n_k \neq 0]\!]\, n_k \frac{N_k}{n_k} \left(\varepsilon^k - \sum_{j>k} [\![n_j \neq 0]\!]\, \varepsilon^j \right) \\ &= \sum_{k\geq 0} [\![n_k \neq 0]\!]\, N_k\, \varepsilon^k - \sum_{j=0}^{k-1} [\![n_k \neq 0, n_j \neq 0]\!]\, N_j\, \varepsilon^k \\ &= \sum_{k\geq 0} [\![n_k \neq 0]\!]\, \varepsilon^k \left(N_k - \sum_{j=0}^{k-1} [\![n_j \neq 0]\!]\, N_j \right) \\ &= [\![n_0 \neq 0]\!]\, \varepsilon^0\, N_0 = \mathbf{1}. \end{aligned}$$

For (c), fix a kappa measure κ over Ω. Then,

$$\begin{aligned} &\sum_{\omega\in\Omega} [\![\kappa(\omega) = 0]\!]\, \|\zeta_\kappa(\omega)\|_\rho \leq 1 + \sum_{j\geq 1} \rho^{-j} \leq 1 + \frac{1}{\rho - 1}, \\ &\sum_{\omega\in\Omega} [\![\kappa(\omega) > 0]\!]\, \|\zeta_\kappa(\omega)\|_\rho \\ &\quad \leq \sum_{k>0} [\![n_k \neq 0]\!]\, N_k \left\| \varepsilon^k - \sum_{j>k} [\![n_j \neq 0]\!]\, \varepsilon^j \right\|_\rho \\ &\quad \leq \sum_{k\geq 1} [\![n_k \neq 0]\!]\, N_k \sum_{j\geq k} [\![n_j \neq 0]\!]\, \rho^{-j} \\ &\quad \leq 2^{|\Omega|} \sum_{k\geq 1} [\![n_k \neq 0]\!] \sum_{j\geq k} [\![n_j \neq 0]\!]\, \rho^{-j} \\ &\quad \leq 2^{|\Omega|} \sum_{k\geq 1} [\![n_k \neq 0]\!] \frac{\rho^{-k}}{1 - \rho^{-1}} \leq \frac{|\Omega|\, 2^{|\Omega|}}{\rho - 1} \end{aligned}$$

where we have used $[\![n_k \neq 0]\!]\, N_k \leq 2^{|\Omega|}$ (exercise). So,

$$\sum_{\omega\in\Omega} \|\zeta_\kappa(\omega)\|_\rho \leq 1 + \frac{1 + |\Omega|\, 2^{|\Omega|}}{\rho - 1} \to 1$$

as $\rho \to \infty$ and independently of κ. □

4 Qualitative MDPs

A Qualitative Markov Decision Process (QMDP) is an MDP in which the quantitative information is given by qualitative probabilities and costs, i.e. a QMDP is characterized by

(QM1) a finite set of states $S = \{1, 2, \ldots, n\}$,

(QM2) a finite set of controls $U(i)$ for each $i \in S$,

(QM3) qualitative transition probabilities $P_{i,u}(j)$ of making a transition to $j \in S$ after applying control $u \in U(i)$ in $i \in S$, and

(QM4) a qualitative cost $g(i, u)$ of applying control $u \in U(i)$ in $i \in S$.

To define the cost associated to policy $\pi = (\mu_1, \mu_2, \ldots)$, consider an N-stage π-*trajectory* starting at state i $\tau = \langle i_k \rangle_{k\geq 0}$ where $i_0 = i$ and $P_{i_k,\mu_k(i_k)}(i_{k+1}) > \mathbf{0}$. Each such trajectory τ has qualitative probability and cost given by

$$P(\tau) \doteq \prod_{k=0}^{N-1} P_{i_k,\mu_k(i_k)}(i_{k+1}), \qquad (11)$$

$$g(\tau) \doteq \sum_{k=0}^{N-1} \alpha^k \, g(i_k, \mu_k(i_k)). \qquad (12)$$

The *infinite horizon expected discounted cost* of applying policy π starting at state i is defined as

$$J_\pi(i) \doteq \lim_{N\to\infty} \sum_\tau P(\tau)\, g(\tau) = \lim_{N\to\infty} E\big[g(\tau)\big] \qquad (13)$$

where the sum is over all N-stage π-trajectories starting at i and the expectation is with respect to the qualitative probability (11) (compare with Eq.(1) for MDPs). In general, the limit in (13) is not always well-defined. However, when all costs $g(i,u)$ are in $\mathcal{Q}$ and $\alpha < 1$, then the limit exists and J_π is well-defined. From now on, we will assume that this is the case. The optimal cost-to-go starting from state i, denoted by $J^*(i)$, is $J^*(i) \doteq \inf_\pi J_\pi(i)$. We would like to prove that J^* is well-defined and that there exists a stationary policy μ^* such that $J^* = J_{\mu^*}$. Unfortunately, such result seems very difficult since $\underline{\mathcal{S}}_\rho$ lacks the least upper bound property of the reals. Thus, we conform ourselves with showing the existence of optimal stationary policies and a method for computing them. That is, we need to show that the partial order $\leq$ in $\underline{\mathcal{S}}_\rho^n$ (where $J \leq J'$ if $J(i) \leq J(i), i = 1 \ldots, n$) has a unique minimum in the set $\{J_\mu : \mu \text{ a stationary policy}\}$. For that, we will show that the qualitative version of the Bellman equation has unique solution. Let T be the DP operator for the qualitative MDP. Then,

Theorem 4 *If $\alpha < 1$, then there exists $\rho \geq 1$ such that T is a contraction mapping.*

Proof: Choose ρ large enough so that

$$\gamma \doteq \max_{i=1\ldots n} \sup_{u\in U(i)} \alpha \sum_{j=1}^n \|P_{i,u}(j)\|_\rho < 1,$$

which exists since $U(i)$ is finite. Let $J, H \in \mathcal{S}_\rho^n$. Then,

$$\begin{aligned}
&\|TJ - TH\| \\
&\quad = \max_{i=1\ldots n} \|(TJ)(i) - (TH)(i)\|_\rho \\
&\quad = \max_{i=1\ldots n} \Bigg\| \left(\min_{u\in U(i)} g(i,a) + \alpha \sum_{j=1}^n P_{i,u}(j) J(j) \right) - \\
&\qquad\qquad \left(\min_{u\in U(i)} g(i,a) + \alpha \sum_{j=1}^n P_{i,u}(j) H(j) \right) \Bigg\|_\rho \\
&\quad \leq \max_{i=1\ldots n} \alpha \sum_{j=1}^n \|P_{i,u^*}(j)\|_\rho \, \|J(j) - H(j)\|_\rho \\
&\quad \leq \|J - H\| \max_{i=1\ldots n} \alpha \sum_{j=1}^n \|P_{i,u^*}(j)\|_\rho \\
&\quad \leq \gamma \|J - H\|
\end{aligned}$$

where u^* is the control that minimizes the minimum term in the second equality. □

Corollary 5 *Assume $g(i,u) \in \mathcal{Q}$ for all $i \in S, u \in U(i)$ and $\alpha < 1$. Then, the qualitative version of the Bellman Equation* (2) *has unique solution J^*. In addition, J^* can be found with Value Iteration, and the policy μ^* greedy with respect to J^* is the best stationary policy.*

Note, however, that μ^* is not necessarily optimal. This possibility seems very unlikely but a definite answer is yet to be found.

4.1 Order-of-Magnitude Specifications

We say that a QMDP is an *order-of-magnitude specification* when the transition probabilities $P_{i,u}(j)$ are only known up to a compatible kappa ranking $\psi_{i,u}(j)$, i.e. $P_{i,u}(j)^\circ = \psi_{i,u}(j)$. In this case, we consider the QMDP the corresponds to the operator:

$$(TJ)(i) = \min_{u\in U(i)} g(i,u) + \alpha \sum_{j=1}^n \zeta_{\psi_{i,u}}(j) J(j)$$

where $\zeta_{\psi_{i,u}}$ is the projection of $\psi_{i,u}$. Clearly, Corollary 5 asserts that this QMDP can be solved with Value Iteration.

5 Qualitative POMDPs

A definition for Qualitative POMDPs (QPOMDP) can be obtained readily from the POMDP formulation:

- (QP1) A finite state space $S = \{1, \ldots, n\}$,
- (QP2) a finite set of controls $U(i)$ for each $i \in S$,
- (QP3) qualitative transition probabilities $p_{i,u}(j)$ of making a transition to j when control $u \in U(i)$ is applied in $i \in S$,
- (QP4) a finite set of observations $O(i,u) \subseteq O$ that may result after applying $u \in U(i)$ in $i \in S$,
- (QP5) qualitative observation probabilities $p_{i,u}(o)$ of receiving observation $o \in O(i,u)$ in $i \in S$ after the application of $u \in U(i)$, and
- (QP6) a qualitative cost $g(i,u)$ associated to $u \in U(i)$ and $i \in S$.

Similarly, we can define a qualitative version of the belief-MDP but a serious problem appears: the infiniteness of the belief-MDP thwarts a suitable choice for ρ as in Theorem 4. Consequently, we are not able to proof that the corresponding DP operator is a contraction. Fortunately, we can get good results for the case of order-of-magnitude specifications.

5.1 Order-of-Magnitude Specifications

A QPOMDP is said to be an *order-of-magnitude specification* if the qualitative probabilities $P_{i,u}(j)$ and

$P_{i,u}(o)$ are only known up to compatible kappa rankings $\psi_{i,u}(j)$ and $\vartheta_{i,u}(o)$ respectively. In this case, we are content to only consider *kappa belief states* which are defined as kappa rankings over states. Thus, the transition dynamics of kappa belief states is given by the order-of-magnitude analogous to Eqs.(6)–(9):

$$\kappa_u^o(i) \doteq \kappa_u(i) + \vartheta_{i,u}(o) - \kappa^u(o), \tag{14}$$

$$\kappa_u(i) \doteq \min_{j=1\ldots n} \Big[\kappa(j) + \psi_{j,u}(i)\Big], \tag{15}$$

$$\kappa^u(o) \doteq \min_{i=1\ldots n} \Big[\kappa_u(i) + \vartheta_{i,u}(o)\Big], \tag{16}$$

$$O(k,u) \doteq \{o : k^u(o) < \infty\}. \tag{17}$$

The qualitative belief-MDP over kappa belief states is:

(K1) A belief space K of kappa measures over S,

(K2) a set of controls $U(\kappa) = \{u : \forall i[\kappa(i) < \infty \Rightarrow u \in U(i)]\}$ for each $\kappa \in K$, and

(K3) a cost $g(k,u) = \sum_{i=1}^n g(i,u)\,\zeta_\kappa(i)$ for each $u \in U(\kappa)$ and $\kappa \in K$ where ζ_κ is the projection of κ.

With transition probabilities given by

$$\kappa \rightsquigarrow \kappa_u^o \quad \text{with qualitative probability} \quad \zeta_{\kappa^u}(o).$$

Therefore, the Bellman equation is

$$J^*(\kappa) = \min_{u\in U(\kappa)} g(\kappa,u) + \alpha \sum_{o\in O(\kappa,u)} \zeta_{\kappa^u}(o) J^*(\kappa_u^o) \tag{18}$$

with the obvious DP operator. As claimed, we now prove that the Bellman equation (18) has unique solution that can be found with Value Iteration.

Theorem 6 *If $\alpha < 1$, then there exists $\rho > 1$ such that the* DP *operator T is a contraction mapping.*

Proof: Choose ρ large enough such that

$$\sup_{\kappa} \sup_{u\in U(\kappa)} \alpha \sum_{o\in O(\kappa,u)} \|\zeta_{\kappa^u}(o)\|_\rho < 1$$

where the first sup is over all kappa rankings over S. The existence of ρ is guaranteed by $\alpha < 1$ and Theorem 3-(c). Then, use a proof similar to that of Theorem 4. □

Corollary 7 *Assume $g(i,u) \in \mathcal{Q}$ for all $i \in S, u \in U(i)$ and $\alpha < 1$. Then, the Bellman Equation* (18) *has unique solution J^*. In addition, J^* can be found with Value Iteration, and the policy μ^* greedy with respect to J^* is the best stationary policy.*

6 Computational Issues

So far, we have focused the paper in the theoretical foundations for a qualitative theory of MDPs and POMDPs. In the computational side, we showed that the Value Iteration algorithm can be used to find optimal stationary policies for QMDPs and QPOMDPs. As in the standard theory, Value Iteration is one of the most general algorithms for solving sequential decision tasks, yet for certain subclasses of problems other algorithms might be more efficient. For example, problems in which the agent does not receive any feedback from the environment (i.e. problems with a single *dummy* observation that is always received) form an important class known as *conformant* planning problems. They can be efficiently solved by performing search in belief space with standard algorithms like A* or IDA*. These search algorithms are guided by heuristic functions that can be automatically extracted from the problem representation. Such methods had proven to be powerful and successful in the standard setting and we expect them to do well in this setting; see [4, 5].

Another important subclass is that in which all transition and observation probabilities has the same order of magnitude 0. In such case, the kappa belief states are just sets of states and the resulting model can be solved using the model-checking or SAT based approaches to planning; see [7, 2, 17].

Finally, in settings in which the number of steps is bounded a priori, the qualitative planning problem can be encoded into a (qualitative) probabilistic SAT formula that can be solved by a suitable modification of the MAXPLAN algorithm [19].

7 Discussion

The work in kappa rankings was first formalized by Spohn [25] but its roots can be traced back to Adam's conditionals [1]. They have been used by Pearl and Goldszmidt to define a qualitative decision theory [22, 21, 14] and also are connected with the ϵ-semantics for default reasoning [20, 12]. More recently, Giang and Shenoy present a qualitative utility theory based in kappa rankings [13].

Another approach to qualitative MDPs and POMDPs had been recently given in terms of possibility theory [11, 24]. This approach is based on the qualitative decision criteria within the framework of possibility theory suggested in [10]. As our approach, theirs computes a value function and policy using a suitable version of the Value Iteration algorithm. However, as their example shows, knowing the (possibilistic) cost function is not sufficient for recovering the policy (this is the reason why their algorithm computes the policy together with the cost function). This is a fundamental departure from the standard theory of MDPs and POMDPs in which there is a one-one correspondence between stationary policies and cost functions. The reason for such discrepancy can be better understood by considering the order-of-magnitude version of the

Bellman equation:

$$J^*(i) = \min_{u \in U(i)} \max\left\{ g(i,u)^\circ, \max_{j=1\dots n} \left(p_{i,u}(j)^\circ + J^*(j) \right) \right\}.$$

It is not hard to see that this equation is (usually) constant for all states except the goal, so it cannot be used to discriminate among controls. This loss of information is due to the fact that order-of-magnitude quantities cannot "increase" through summations (a fact that is well-known to researchers in the field). This is the main reason why we have decided to work with the Extended Reals instead of their orders of magnitude.

Another important point we need to discuss is that the choice for projections seems quite arbitrary. This fact, however, does not worry us since the kappa specification is very imprecise and so the only important thing that is required when going from kappa rankings to qualitative probabilities is consistency. Thus, we would feel as happy with any other consistent map that can do the job.

In summary, we have proposed a novel formulation for a qualitative theory of MDPs and POMDPs that is based upon a complete extension of the Extended Reals proposed by Wilson. The formal developments had been achieved using Mathematical ideas from Functional Analysis, and was motivated by the necessity of taking limits. The new entities are then combined using standard techniques in the theory of MDPs to obtain a qualitative theory that is very close to the standard theory. This is an important difference with other approaches whose ties with the standard theory of MDPs and POMDPs are not as clear.

References

[1] E. Adams. *The Logic of Conditionals*, chapter 2. D. Reidel, Dordrecht, Netherlands, 1975.

[2] P. Bertoli, A. Cimatti, and M. Roveri. Heuristic search + symbolic model checking = efficient conformant planning. In *Proceedings of IJCAI-2001*, pages 467–472, 2001.

[3] D. Bertsekas. *Dynamic Programming and Optimal Control, (2 Vols)*. Athena Scientific, 1995.

[4] B. Bonet and H. Geffner. Planning with incomplete information as heuristic search in belief space. In *Proceedings of AIPS-2000*, pages 52–61. AAAI Press, 2000.

[5] B. Bonet and H. Geffner. Gpt: a tool for planning with uncertainty and partial information. In *Proceedings of the IJCAI-01 Workshop on Planning with Uncertainty and Partial Information*, pages 82–87, 2001.

[6] A. Cassandra, L. Kaebling, and M. Littman. Acting optimally in partially observable stochastic domains. In *Proceedings AAAI94*, pages 1023–1028, 1994.

[7] A. Cimatti and M. Roveri. Conformant planning via symbolic model checking. *Journal of Artificial Intelligence Research*, 13:305–338, 2000.

[8] T. Dean, L. Kaebling, J. Kirman, and A. Nicholson. Planning with deadlines in stochastic domains. In *Proceedings AAAI93*, pages 574–579. MIT Press, 1993.

[9] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing and Uncertainty.* Plenum Press, New York, 1988.

[10] D. Dubois and H. Prade. Possibility theory as a basis for qualitative decision theory. In *Proceedings of IJCAI-95*, pages 1925–1930, 1995.

[11] H. Fargier, J. Lang, and R. Sabbadin. Towards qualitative approaches to multi-stage decision making. *Int. Journal of Approximate Reasoning*, 19:441–471, 1998.

[12] H. Geffner. *Default Reasoning: Causal and Conditional Theories.* The MIT Press, 1992.

[13] P. Giang and P. Shenoy. A qualitative linear utility theory for spohn'theory of epistemic beliefs. In *Proceedings UAI-2000*, pages 220–229. Morgan Kaufmann, 2000.

[14] M. Goldszmidt and J. Pearl. System Z^+: A formalization for reasoning with variable strength defaults. In *Proceedings of AAAI-91*, pages 399–404, 1991.

[15] M. Hauskrecht. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

[16] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1999.

[17] H. Kautz and B. Selman. Pushing the envelope: Planning, propositional logic, and stochastic search. In *Proceedings of AAAI-96*, pages 1194–1201, Protland, Oregon, 1996. MIT Press.

[18] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis.* Dover Publications, 1975.

[19] S. Majercik and M. Littman. Maxplan: A new approach to probabilistic planning. In *Proceedings of Fourth International Conference on Artificial Intelligence Planning and Schedulin*, pages 86–93, 1998.

[20] J. Pearl. Epsilon semantics. In *Encyclopedia of AI*, pages 468–475. John Wiley & Sons, Inc., 2nd edition, 1992.

[21] J. Pearl. From conditional oughts to qualitative decision theory. In *Proceedings UAI-93*, 1993.

[22] J. Pearl. From adams' conditionals to default expressions, causal conditionals, and counterfactual. In E. Eells and B. Skyrms, editors, *Probability and Conditionals.* Cambridge University Press, 1994.

[23] H. L. Royden. *Real Analysis.* Collier MacMillan, second edition, 1968.

[24] R. Sabbadin. A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments. In *Proceedings UAI-99.* Morgan Kaufmann, 1999.

[25] W. Spohn. A general non-probabilistic theory of inductive reasoning. In *Proceedings 4th Workshop on Uncertainty*, pages 315–322, St. Paul, 1988.

[26] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[27] N. Wilson. An order of magnitude calculus. In *Proceedings of UAI-95*. Morgan Kaufmann, 1995.