HAL
open science

Abstraction Pathologies In Markov Decision Processes

Manel Tagorti, Bruno Scherrer, Olivier Buffet, Joerg Hoffmann

To cite this version:

Manel Tagorti, Bruno Scherrer, Olivier Buffet, Joerg Hoffmann. Abstraction Pathologies In Markov Decision Processes. 8èmes Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes, Jul 2013, Lille, France. hal-00907295

HAL Id: hal-00907295

https://inria.hal.science/hal-00907295v1

Submitted on 21 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Abstraction Pathologies In Markov Decision Processes

Manel Tagorti[1], Bruno Scherrer[1], Olivier Buffet[1], Jörg Hoffmann[2]

[1] INRIA
615, rue du Jardin Botanique 54600 Villers-lès-Nancy.
{manel.tagorti, bruno.scherrer, olivier.buffet}@inria.fr
[2] Saarland University, Saarbrücken, Germany
hoffmann@cs.uni-saarland.de

Abstract : Abstraction is a common method to compute lower bounds in classical planning, imposing an equivalence relation on the state space and deriving the lower bound from the quotient system. It is a trivial and well-known fact that refined abstractions can only improve the lower bound. Thus, when we embarked on applying the same technique in the probabilistic setting, our firm belief was to find the same behavior there. We were wrong. Indeed, there are cases where *every* direct refinement step (splitting one equivalence class into two) yields strictly *worse* bounds. We give a comprehensive account of the issues involved, for two wide-spread methods to define and use abstract MDPs.
Keywords : Abstractions, Markov Decision Processes, Heuristic.

1 Introduction

In classical planning, an abstraction is a mapping α from the set of all states into a smaller set of abstract states ($[s]_\alpha$ denoting the set of states t where $\alpha(t) = \alpha(s)$). This is used to derive a lower bound $h^\alpha(s)$ on the remaining cost of any state s. Namely, α induces an abstract planning problem over the abstract state space: (i) an abstract state $[s]_\alpha$ is a goal iff it contains at least one original goal state, and (ii) a transition from $[s]_\alpha$ to $[s']_\alpha$ exists iff there exist $t \in [s]_\alpha$ and $t' \in [s']_\alpha$ so that the original state space has a transition from t to t'. The abstract planning problem is a relaxed version of the original one –or, conversely, the original problem is more constrained– so that, given a state s, the cost of an abstract plan starting from $[s]_\alpha$ is at most equal to the cost of a plan starting from s, and can thus be used as a lower bound $h^\alpha(s)$. Prominent examples of this method are pattern databases Edelkamp (2001); Haslum *et al.* (2007) and merge-and-shrink abstractions Helmert *et al.* (2007); Nissim *et al.* (2011); Katz *et al.* (2012).

A refinement of α is an abstraction α' resulting from α by splitting some of the block states, i.e., for all s we have $[s]_{\alpha'} \subseteq [s]_\alpha$. It is commonplace that refinements can only improve the heuristic, i.e., $h^\alpha(s) \leq h^{\alpha'}(s)$ for all s: If we split block states apart, then the solution paths can only get longer and thus more costly (assuming non-negative costs as usual). Indeed, this observation is so simple that, to our knowledge, no-one yet bothered to state it in a paper and its first appearance is in Malte Helmert's 2010 lecture slides.[1]

Our initial agenda in this research was to solve MDPs using heuristic search methods like LRTDP Bonet & Geffner (2003), our focus being to compute heuristic functions by starting with a coarse abstraction and iteratively refining it. Against the background described above, as a warm-up exercise we embarked on proving that the essential property of refinements – they can only improve the heuristic – is true in that setting as well. Which was all fine, except we ended up proving the opposite.

First things first, to conduct this kind of research for MDPs one needs to first define what the "quotient system" and the corresponding heuristic functions are. This is non-trivial because, in difference to the classical case where all we are interested in is which states can transition to which other states in principle, now we need to define transition *probabilities* for the abstract MDP. To illustrate, if action a maps s into a state from $[s']_\alpha$ with probability 0.9, but maps $t \in [s]_\alpha$ into a state from $[s']_\alpha$ with probability 0.1, which probability should we assign for a to map $[s]_\alpha$ into $[s']_\alpha$?

A simple answer is to assign the average probability over all states in $[s]_\alpha$. In the example, this would yield the transition probability 0.5. A main issue with this approach is that the resulting heuristic function

[1] http://www.informatik.uni-freiburg.de/~ki/teaching/ss10/aip/aip10.pdf

– the value function of the abstract MDP – is neither a lower bound nor an upper bound on the value function of the original MDP. Givan et al. Givan *et al.* (2000) fix this by basically considering intervals of transition probabilities (in the example, the interval $[0.1, 0.9]$). They derive a lower bound on the original value function (expected reward) by selecting the probabilities pessimistically, and derive an upper bound of the original value function by selecting the probabilities optimistically.

We first proved that, for the average-probability approach, there exist an MDP, a state s, an abstraction α, and refined α' so that the error of $h^{\alpha'}(s)$ relative to the original value function is larger than that of $h^{\alpha}(s)$. This may be duly understood as an accident pertaining to the sketchy nature of this approach; indeed, as we show, the original MDP does not have to be non-deterministic to provoke this kind of behavior. However, we next proceeded to prove the same property for Givan et al.'s approach. Worse, even: We constructed an MDP, a state s, and α so that *all* direct refinements α' (resulting from α by splitting a single block state) result in strictly worse bounds. More naturally than for the average-probability approach, non-determinism is required for this, i.e., if the original MDP is deterministic then every refinement results in better bounds.

In the remainder of the paper, we first give the necessary background definitions. We then explain our results in the order described above.

2 Markov Decision Processes and Abstractions

Markov Decision Processes (MDPs) are a general framework for modeling decision-making problems in stochastic environments. We define an MDP as follows

Definition 1. *A Markov Decision Process is given by a (finite) state space S, a (finite) action space A, a reward function $R : S \times A \to \mathbb{R}$ and transition probabilities $p(s, a, s')$ which determine the probabilities of transition when performing action a in state s.*

A deterministic policy $\pi : S \to A$ assigns an action to each state, and we are looking for the optimal policy π^*, i.e., one that maximizes for all $s \in S$ the return

$$V^{\pi}(s) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t | s_0 = s\right],$$

where γ is the discount rate taken in $[0, 1)$.

The related value V^* is the unique solution of the equation $V^* = TV^*$ where T is the Bellman operator defined as :

$$\forall s,\ V^*(s) = TV^*(s) = \max_a \left(R(s, a) + \gamma \sum_{s' \in S} p(s, a, s') V^*(s') \right).$$

Determining π^* may be infeasible in MDPs with large state spaces. In this paper we simplify the problem by employing state abstractions. Abstractions provide a smaller representation M_α of the original MDP M. The image of M under an abstraction α is defined on a state space S_α smaller than S. Indeed S_α is a partition of S consisting of block states $[s]_\alpha$. We assign to each block-state, given an action $a \in A$, the reward $R([s]_\alpha, a)$ and the transition probabilities $p([s]_\alpha, a, [s_1]_\alpha)$ for all $[s_1]_\alpha \in S_\alpha$. The useful abstractions are the ones that induce a small error of approximation when considering M_α instead of M. We would like to identify such abstractions by comparing a given abstraction to its (direct) refinement, in terms of approximation error.

2.1 Abstractions' Refinement

Definition 2. *Let α and α' be two abstractions of an MDP M. We say that α' is* finer *than α, denoted $\alpha' \succeq \alpha$, iff for any states $s, s' \in S$, $\alpha'(s) = \alpha'(s')$ implies $\alpha(s) = \alpha(s')$. We can also say that α is* coarser *than α', denoted $\alpha \preceq \alpha'$.*

We have α' a direct refinement *of α if there exist states $s_1, s_2 \in S$ such that $[s_1]_\alpha = [s_2]_\alpha$, $[s_1]_{\alpha'} \neq [s_2]_{\alpha'}$, $[s_1]_\alpha = [s_1]_{\alpha'} \cup [s_2]_{\alpha'}$, and $\alpha'(s) = \alpha(s)$ for all $s \in S \setminus [s_1]_\alpha$.*

We show in what follows that the error induced by α' may in some cases be higher than the one induced by α. But before that we have first to specify the parameters (rewards and transition probabilities) related to the abstract representation M_α of M.

3 Average MDPs

We consider in this section the abstraction α that connects an MDP M to its average representation. In other words, α maps an MDP M defined on S to an average MDP M_α, defined on S_α, and admits as parameters, the averages of rewards and transitions over all states contained in the block state $[s]_\alpha$ (Ortner, 2011), i.e., we have for all $a \in A$:

$$R([s]_\alpha, a) = \frac{1}{|[s]_\alpha|} \sum_{s_1 \in [s]_\alpha} R(s_1, a) \text{ and}$$

$$p([s]_\alpha, a, [s']_\alpha) = \frac{1}{|[s]_\alpha|} \sum_{s_1 \in [s]_\alpha} \sum_{s_2 \in [s']_\alpha} p(s_1, a, s_2).$$

We denote here by $|[s]_\alpha|$ the cardinal of all states in $[s]_\alpha$.

We choose as approximation error E_α, the average error, estimated by taking the average difference between the true value V taken in a state s and the value V_α of its corresponding block state $[s]_\alpha$,

$$E_\alpha = \frac{1}{|S|} \sum_{s \in S} |V_\alpha([s]_\alpha) - V^*(s)|.$$

This approximation error may increase when we refine the abstraction α. We illustrate in Figure 1 an example in which the number of states where the (local) error increases-after a refinement-is greater than the one where the (local) error decreases, causing the increase of the average error.

Proposition 1. *There exists a deterministic MDP M, an abstraction α and a refinement α' of α such that $E_\alpha < E_{\alpha'}$.*

Proof. Consider the MDP M in Figure 1 with a single action $\{a\}$ and a discount rate $\gamma = 1$ (the result would not change for $\gamma < 1$ but close enough to 1). The states in $\{2, ..., k\}$ ($k > 2$) are similar: they admit the same rewards ($R = 0$) and have the same dynamics : they reach the neighboring state with probability one. The states in $\{k+1, ..., n\}$ are also similar: they all reach the goal G with probability one and they admit a non-negative reward R_2. The state 1 admits a reward $R(1) = R_1 \ll R_2$ and reaches the goal with probability one.

Taking $V(G) = 0$, we then have $V(1) = R_1$, and $V(i) = R_2$ forall i in $\{2, ..., n\}$.

Based on those similarities one can construct a perfectly suitable abstraction α_0 ($E_{\alpha_0} = 0$) which aggregates similar states in the same block, i.e., in our case $\alpha_0 : S \to 1, \{2, ..., k\}, \{k+1, ..., n\}, G$.

Consider now the abstraction $\alpha_1 : S \to \{1, ..., k\}, \{k+1, ..., n\}, G$, where the states 1 and $\{2, ..., k\}$ are in the same block (Figure 1). The similarity will be then broken resulting in a strictly positive error E_α. The related values are

$$V_{\alpha_1}(\{k+1, ..., n\}) = \frac{(n-k)R_2}{n-k} = R_2 \text{ and}$$

$$V_{\alpha_1}(\{1, ..., k\}) = \frac{R_1}{k} + \frac{k-2}{k} V_{\alpha_1}(\{1, ..., k\}) + \frac{1}{k} V_{\alpha_1}(\{k+1, ..., n\}) = \frac{R_1 + R_2}{2} \neq R_2.$$

And the induced error is

$$E_{\alpha_1} = \frac{1}{n} \sum_{i=1}^{k} |V(i) - V_{\alpha_1}(\{1, ..., k\})| \sim k \frac{R_2}{2n} \text{ for } R_1 \ll R_2.$$

Let $\alpha_2 : S \to \{1, ..., n\}, G$ be the abstraction which aggregates states $1, 2, ..., n$ together (Figure 1), the value $V_{\alpha_2}(\{1, ..., n\})$ is equal to

$$V_{\alpha_2}(\{1, ..., n\}) = \frac{R_1 + (n-k)R_2}{n} + \frac{k-1}{n} V_{\alpha_2}(\{1, ..., n\}) \sim R_2 \text{ for } k \ll n$$

and hence

$$E_{\alpha_2} \sim \frac{1}{n} |V(1) - V_{\alpha_2}(\{1, ..., n\})| \sim \frac{1}{n} R_2.$$

We can see that the error E_{α_1} is strictly larger than E_{α_2} for a number of states k strictly greater than 2. □

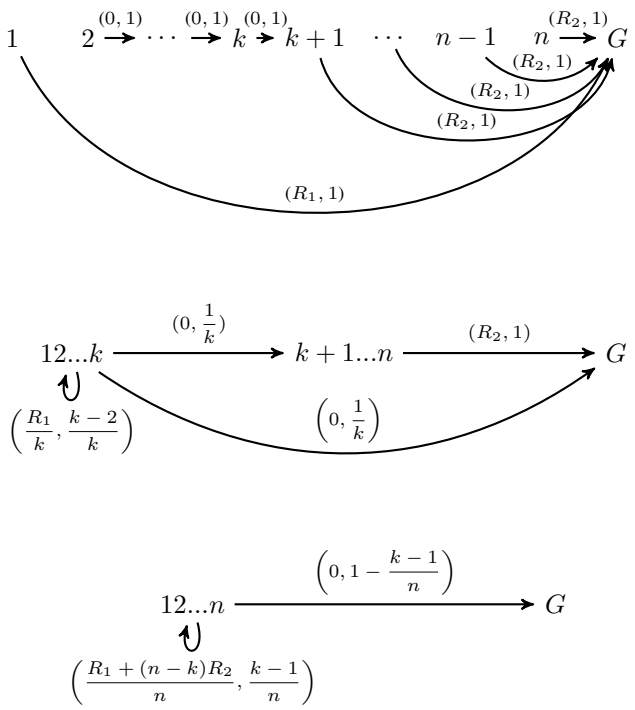

Figure 1: From the top to the bottom: The MDP M, the abstraction α_1 and the abstraction α_2 (α_0 is not shown). The parentheses denote the reward (on the left) and the transition probabilities (on the right).

4 Bounded-parameter MDPs

Rather than approaching V^* with fixed values V_α, it is possible to establish bounds on the MDP's parameters so as to get bounds on V^*. So we considered the abstraction α that associates, for each (action) a in A, each block state $[s]_\alpha$ with the intervals' parameters ((Givan *et al.*, 2000)):

$$R^{\updownarrow}([s]_\alpha, a) = [\min_{s\in[s]_\alpha} R(s,a), \max_{s\in[s]_\alpha} R(s,a)]$$
$$p^{\updownarrow}([s]_\alpha, a, [s']_\alpha) = [\min_{s_1\in[s]_\alpha} \sum_{s_2\in[s']_\alpha} p(s_1,a,s_2), \max_{s_1\in[s]_\alpha} \sum_{s_2\in[s']_\alpha} p(s_1,a,s_2)].$$

We get hence what we call a *a bounded parameter Markov Decision process* (BMDP) more commonly defined as:

Definition 3. *A Bounded parameter Markov Decision Process is given by a (finite) state space Σ, a (finite) action space A, an interval of rewards $R^{\updownarrow}(\sigma, a)$, $\forall \sigma \in \Sigma, a \in A$, and an interval of transition probabilities $p^{\updownarrow}(\sigma, a, \sigma')$, $\forall \sigma, \sigma' \in \Sigma, a \in A$.*

Each state of the BMDP has a range of values depending on R and p. We can assign to each state a closed interval of value functions $[V^-(\sigma), V^+(\sigma)]$ where V^- corresponds to the pessimistic bound and V^+ to the optimistic one.

Our abstract representation M_α is then a BMDP where the state space Σ coincides with state space S_α. We would like to estimate the value bounds V_α^+ ans V_α^- related to M_α. Givan *et al.* have proposed an algorithm, the interval value iteration IVI, to do so. To make explicit the two Bellman's operators hidden behind this algorithm (T^+ and T^-), we first need to introduce the notion of *compatibility* with respect to an abstraction.

Definition 4. *An MDP N is compatible with M with respect to the abstraction α if for all s, and for all a, $R_N(s,a) \in R_M^{\updownarrow}([s]_\alpha, a)$ and for all s, s', a, $\sum_{s_1\in[s']_\alpha} p_N(s,a,s_1) \in p_M^{\updownarrow}([s]_\alpha, a, [s']_\alpha)$.*
The set of all MDPs compatible with M with respect to α is denoted $[M]_\alpha$.

Figure 2 gives an example of an MDP N compatible with an MDP M with respect to the abstraction $\alpha : 1, 2, 3 \to \{1, 2\}, 3$.

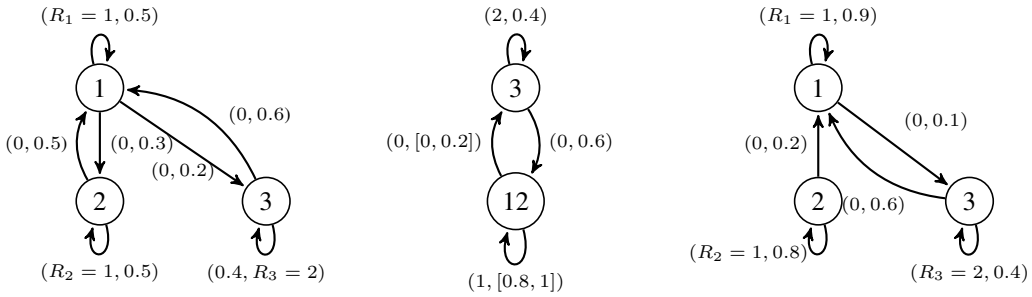

Figure 2: From left to right, the MDP M, the BMDP M_α, $\alpha : 1, 2, 3 \to \{1, 2\}, 3$ and the MDP N compatible with M with respect to α.The parenthesis denote the reward (on the left) and the transitions (on the right)

We claim that the Bellman operators T^+ and T^- used to estimate V_α^+ and V_α^- can be written in this way, for all s:

$$T^+[V](s) = \max_{a \in A} \max_{N \in [M]_\alpha} R_N(s, a) + \gamma \sum_{s' \in S} p_N(s, a, s')V(s'),$$
$$T^-[V](s) = \max_{a \in A} \min_{N \in [M]_\alpha} R_N(s, a) + \gamma \sum_{s' \in S} p_N(s, a, s')V(s').$$

By taking iteratively the $\max$ (resp the $\min$) on the set of compatible MDPs and by choosing the optimal policy, we can see that those two operators converge to fixed values. Indeed T^+ and T^- are γ-contracting so, by the Banach fixed point Theorem, they admit unique fixed points V_α^+ and V_α^- (which are constant per block), i.e., $T^+V_\alpha^+ = V_\alpha^+$ and $T^-V_\alpha^- = V_\alpha^-$. There exists an optimistic MDP M^{opt} (respectively a pessimistic MDP M^{pes}) and a corresponding optimal (optimistic) policy π^{opt} (respectively an optimal (pessimistic) policy π^{pes}) for which the value V_α^+ (resp V_α^-) is reached. The MDPs M^{opt} and M^{pes} belong to $[M]_\alpha$.

Before proceeding to the next step –the choice of the abstraction that would ensure better bounds–, we first show that these values are indeed bounds on V^*. This is precisely what is stated in this following theorem.

Theorem 1. *(Givan* et al.*, 1997) For any MDP M and abstraction α of the states of M, bounds on the BMDP M_α apply also to M, i.e., $\forall s \in S$, $V^*(s) \in [V_\alpha^-([s]_\alpha), V_\alpha^+([s]_\alpha)]$.*

Proof. [2] The proof is done using Value Iteration. With initial value $V^0 = V_\alpha^+$ we have, for all s,

$$V^1(s) = \max_a R_M(s, a) + \gamma \sum_{s' \in S} p_M(s, a, s')V_\alpha^+([s']_\alpha).$$

Since V_α^+ is a fixed point of T^+ and $M \in [M]_\alpha$, we can see that

$$\begin{aligned} V^1(s) &\leq \max_{a \in A} \max_{N \in [M]_\alpha} R_N(s, a) + \gamma \sum_{s' \in S} p_N(s, a, s')V_\alpha^+([s']_\alpha) \\ &\leq V_\alpha^+([s]_\alpha) = V^0(s). \end{aligned}$$

The Bellman operator T is monotone, so that we have $T^nV^1 \leq T^nV^0 = V^0$. By taking the limit, we get $V^*(s) \leq V_\alpha^+([s]_\alpha)$. A similar proof may be applied to the pessimistic bound. □

[2] *This proof does not appear in (Givan* et al.*, 1997) but it has been established thanks to a correspondence with R. Givan.*

4.1 Value Bounds using finer abstractions

As previously done for the average model, we will compare the two errors G_α and $G_{\alpha'}$ induced by the BMDP M_α and its direct refinement $M_{\alpha'}$, where, for an abstraction α, G_α measures the gap between the two bounds in each state, for all s:

$$G_\alpha(s) = \left(V_\alpha^+([s]_\alpha) - V_\alpha^-([s]_\alpha)\right).$$

The proposition below states a sufficient condition under which the finer abstraction α' yields better value bounds than α and therefore decreases the error G_α. Indeed, if the set of MDPs compatible with M with respect to the finer abstraction are also compatible with M with respect to the coarser abstraction, then we get the inclusion of the value function intervals.

Proposition 2. *Given an MDP M and two abstractions α and α' s.t. $\alpha' \succeq \alpha$ and $[M]_{\alpha'} \subseteq [M]_\alpha$ then, for all $s \in S$,*

$$[V_{\alpha'}^-([s]_{\alpha'}), V_{\alpha'}^+([s]_{\alpha'})] \subseteq [V_\alpha^-([s]_\alpha), V_\alpha^+([s]_\alpha)].$$

Proof. Value Bounds computation can be considered as a case of an alternating two players stochastic game where one choose the optimal policy while the other choose the optimal MDP. For the upper bound, we can then invert the two $\max$ in the Bellman operators' expressions ((Givan *et al.*, 2000)). This would not change the final result, so we have

$$V_\alpha^+([s]_\alpha) = \max_{N \in [M]_\alpha} V_N^*(s).$$

For the lower bound, more detailed arguments have been established in Bertsekas & Tsitsiklis (1996) to set the inversion of the $\max$ and the $\min$ terms, so we get:

$$V_\alpha^-([s]_\alpha) = \min_{N \in [M]_\alpha} V_N^*(s).$$

Given that $[M]_{\alpha'}$ is included in $[M]_\alpha$, then by taking the $\max$, (respectively the $\min$) the result follows. □

We will study next two cases : the deterministic case, where the sufficient condition is always satisfied, and the stochastic case, where it is not necessarily satisfied, for which we will give an example.

4.1.1 Deterministic case: probabilities in $\{0, 1\}$

Corollary 1. *If we consider a deterministic MDP and two abstractions α and α', where α' is a direct refinement of α, then for all $s \in S$,*

$$[V_{\alpha'}^-([s]_{\alpha'}), V_{\alpha'}^+([s]_{\alpha'})] \subseteq [V_\alpha^-([s]_\alpha), V_\alpha^+([s]_\alpha)].$$

Proof. Let us consider an MDP N compatible with M under α'. We will show that the sufficient condition of Proposition 2 is satisfied: N is also compatible with M under α. Note that having N compatible with M with respect to an abstraction α, according to Definition 4, is equivalent to having the inclusion of the parameter intervals i.e., for all s, and for all a, $R_N^\updownarrow([s]_\alpha, a) \subseteq R_M^\updownarrow([s]_\alpha, a)$ and for all s, s', a, $p_N^\updownarrow([s]_\alpha, a, [s']_\alpha) \subseteq p_M^\updownarrow([s]_\alpha, a, [s']_\alpha)$. Let $[s_1]_\alpha$ be the state block in S_α that we split into two blocks $[s_2]_{\alpha'}$ and $[s_3]_{\alpha'}$ (i.e., we have $[s_2]_{\alpha'} \cup [s_3]_{\alpha'} = [s_1]_\alpha$). It is easy to check the inclusion of reward intervals under α as for each action a:

$$\min_{s \in [s_1]_\alpha} R_N(s, a) = \min(\min_{s \in [s_2]_{\alpha'}} R_N(s, a), \min_{s \in [s_3]_{\alpha'}} R_N(s, a)).$$

Also, using the α' compatibility hypothesis we have

$$\min_{s \in [s_i]_{\alpha'}} R_N(s, a) \geq \min_{s \in [s_i]_{\alpha'}} R_M(s, a) \text{ for } i \text{ in } \{2, 3\}$$

then

$$\min_{s \in [s_1]_\alpha} R_N(s, a) \geq \min_{s \in [s_1]_\alpha} R_M(s, a).$$

By reasoning in a similar way for the upper bound, we get the inclusion of the reward intervals. The same arguments may be employed to state the inclusion of outgoing transition probabilities $p([s_1]_\alpha, a, [s_4]_\alpha)$ for all s_4 in S and a in A. So we will mainly focus on ingoing transition probabilities $p([s_5]_\alpha, a, [s_1]_\alpha)$ for all s_5 in S and we will show that

$$\min p_N([s_5]_\alpha, a, [s_1]_\alpha) \geq \min p_M([s_5]_\alpha, a, [s_1]_\alpha). \tag{1}$$

Since we work in a deterministic environment, probabilities can only take the values 0 or 1. For the case where $\min p_N([s_5]_\alpha, a, [s_1]_\alpha) = 1$, inequality (1) is always verified. Now, $\min p_N([s_5]_\alpha, a, [s_1]_\alpha) = 0$ implies that there exist a state s' in $[s_5]_\alpha$ and a block state $[s_6]_\alpha$ distinct from $[s_1]_\alpha$ such that $p_N(s', a, [s_6]_\alpha) = 1$. So we can find a state s'' in $[s_5]_\alpha$ such that $p_M(s'', a, [s_6]_\alpha) = 1$ as N is compatible with M under α'. We then have Equation (1) and by Proposition 2 the final result follows.

□

4.1.2 Stochastic case

We would like to have an equivalent of Proposition 2 for the stochastic case but it turns out that in general the sufficient condition is no more fulfilled. In fact when we move to stochastic transitions, the sufficient condition in Proposition 2 becomes harder to satisfy. The successful MDP M has to verify specific conditions that depend on the choice of the abstractions α and α'. In other words, given an MDP M and an abstraction α there does not always exist a refined abstraction α' such that the sufficient condition is satisfied. Figure 3 shows a model of MDP in which we can not find the appropriate direct refinement α'.Three states which admit the same rewards $R(1) = R(2) = R(3) = 1$ and behave identically in the block $\{1, 2, 3\}$ (the same probabilities in regards to the block $\{1, 2, 3\}$, $p(1, \{1, 2, 3\}) = p(2, \{1, 2, 3\}) = p(3, \{1, 2, 3\}) = 0.7$). States 4 and 5 are goal states ($V(4) = V(5) = 0$). We can find an MDP N compatible with M with respect to the abstraction α' ($S \to \{1, 2\}, 3, 4, 5$) but not compatible with M with respect to the abstraction α ($S \to \{1, 2, 3\}, 4, 5$). It suffices to take an MDP N whose parameter intervals are included in M's parameter intervals under α' but admits real intervals rather than fixed real values under α.

The condition stated in Proposition 2 is a sufficient condition, we can not *a priori* conclude about the existence of a refinement that would make the error of approximation decrease. Nevertheless, we have identified models of MDPs where every direct refinement strictly increases the error.

Proposition 3. *There exists an MDP M and an abstraction α such that, for any direct refinement α' of α, we have: (1) for all s in S, $G_{\alpha'}(s) \geq G_\alpha(s)$, and (2) there exists s in S where $G_{\alpha'}(s) > G_\alpha(s)$.*

Proof. The MDP shown in Figure 3 is an example of such a model. The value $V_\alpha(\{1, 2, 3\})$ related to the block $\{1, 2, 3\}$ is a perfect heuristic for the states $1, 2$ and 3. In fact we can see (for $\gamma = 1$), using Theorem 1, that the optimistic and pessimistic bounds coincide: $V^-_\alpha(\{1, 2, 3\}) = V^+_\alpha(\{1, 2, 3\}) = V(1) = V(2) = V(3) = 3.33$. If we refine the block $\{1, 2, 3\}$ by splitting the block $\{1, 2, 3\}$ into $\{1, 2\}$ and $\{3\}$, then the error of approximation will strictly increase and we will obtain an interval of values rather than a precise value given by the coarser abstraction (we got those values by using Givan et al.'s algorithm). The approximation error G_α equals to 0 for all the states while the approximation error $G_{\alpha'}$ is:

$$G_{\alpha'}(1) = G_{\alpha'}(2) \simeq 1 > 0 = G_\alpha(1) = G_\alpha(2) \text{ and}$$
$$G_{\alpha'}(3) \simeq 1 > 0 = G_\alpha(3).$$

Only one of all the possible refinements is detailed here but the same happens for the two other direct refinements:

- For the abstraction
$$\alpha' : 1, 2, 3, 4, 5 \to \{1, 3\}, \{2\}, \{4\}, \{5\}$$
we have $V^\pm_{\alpha'}(\{1, 3\}) = [2.5, 4.78]$ and $V^\pm_{\alpha'}(\{2\}) = [2.75, 4.34]$.
- For the abstraction
$$\alpha' : 1, 2, 3, 4, 5 \to \{1\}, \{2, 3\}, \{4\}, \{5\}$$
we have $V^\pm_{\alpha'}(\{2, 3\}) = [2.25, 5.29]$ and $V^\pm_{\alpha'}(\{1\}) = [2.9, 4.11]$.

□

We can even have a model inspired from the one above where the error strictly increases in each state. Indeed by taking $\gamma = 0.9$ and by changing the transition probabilities in the initial model M to $p(4,4) = p(5,5) = 0.9$ and $p(4,1) = p(5,1) = 0.1$ (we keep the same rewards $R(4) = R(5) = 0$), we can see that the gap $G_{\alpha'}$ is strictly higher than $G_\alpha = 0$ for each direct refinement α'.

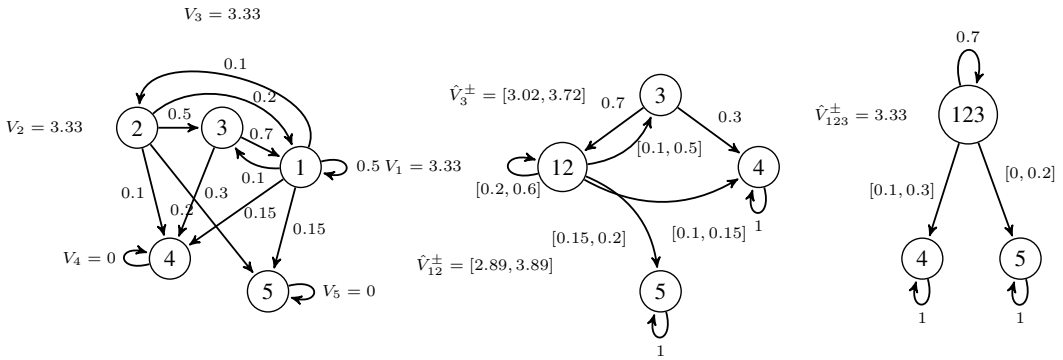

Figure 3: From left to right: the MDP M, the abstraction α', the abstraction α.Edges are annotated with transition probabilities. The reward is: $R(1) = R(2) = R(3) = 1 (= R(\{1,2\}) = R(\{1,2,3\}))$, and $R(4) = R(5) = 0$.

5 Related Work

Our work shows the limitations of some abstractions (state abstractions) in which refining an abstraction may increase the approximation error. This has already been observed in Waugh *et al.* (2009) in the case of an i-player poker game (i greater than 2). They looked at the exploitability[3] of each player's strategy with respect to each abstraction and they showed that it may increase while considering finer card abstractions. The same phenomenon has been observed when they considered betting abstractions – by restricting the number of betting options in each sequence of the game. Lately another example about "action abstractions" pathologies has been provided in Sandholm & Singh (2012).

It is important to notice here that, contrary to what has been done in those works, this paper deals with single action and player models. This suggests that, in some abstractions, notably BMDP abstraction, stochasticity alone can explain this pathological behavior, and that we do not need to consider the more general case of two-player game: the issues appear even in the case of a one-player game (MDP). Indeed, our counter-examples do not even contain any actual action choice, thus identifying this kind of pathology in a very canonical framework.

Interestingly, Kattenbelt et al. Kattenbelt *et al.* (2010) introduce a variant of abstraction for MDPs that, according to their results, does not exhibit said pathology: Every abstraction refinement step results in an improved bound. An interesting open question is how exactly their framework relates to BMDP abstraction.

6 Conclusions

Somewhat surprisingly, refining an abstraction does not guarantee, in the MDP setting, a refined, i.e., better, approximation of the value function. From a practical perspective, this observation might be reasonably classified as "odd but not crucial" – the loss of this guarantee is not per se an argument against trying to apply abstraction techniques for the computation of heuristic functions, as known from classical planning. The observation *might* be relevant to the practical effectiveness of such methods, where paying a higher price for the abstraction may result in less accuracy. But it remains to be seen whether that is of practical importance.

[3]The exploitability is a metric connected to the Nash equilibrium strategy. It is equal to 0 in the case of a two-player extensive game.

From a theoretical perspective, we believe that our observations could be of importance for a better understanding of the methods involved. In that regard, our investigation is but a small start into the subject matter. Does there exist a non-trivial method (not refining all the way to the original MDP) that guarantees, for any MDP and abstraction thereof, the existence of a refinement step reducing the error? And can that method be made practical? In particular, our vision was and is to identify sufficient criteria, in the bounded-parameter MDP setting, for the error to not increase. If such a criterion is efficiently testable, or can at least be reasonably well approximated, then it could serve as a well-informed guidance during the abstraction refinement process. For the moment, we don't know how such a criterion could be designed. An interesting observation in this context is that our counter-example refines an abstraction that already is a bisimulation.[4] Does that tell us something about the general case? Another question is whether increasing the "extent" of the per-step refinement helps (instead of splitting a single block-state, split 2, 3, ... block-states). We believe these are interesting questions for future research, and hope other researchers might join us in exploring them.

References

BERTSEKAS D. P. & TSITSIKLIS J. (1996). In *Neuro-Dynamic Programming*, chapter 7, p. 408–412. Athena Scientific.

BONET B. & GEFFNER H. (2003). Labeled RTDP: Improving the convergence of real-time dynamic programming. In *Proceedings of the 13th International Conference on Automated Planning and Scheduling (ICAPS-03)*, p. 12–21.

EDELKAMP S. (2001). Planning with pattern databases. In *Recent Advances in AI Planning. 6th European Conference on Planning (ECP-01)*, p. 13–24.

GIVAN R., LEACH S. M. & DEAN T. (1997). Model reduction techniques for computing approximately optimal solutions for Markov Decision Processes. *ai*, **122**(1-2), 1–8.

GIVAN R., LEACH S. M. & DEAN T. (2000). Bounded-parameter Markov Decision Processes. *Artificial Intelligence*, **122**(1-2), 71–109.

HASLUM P., BOTEA A., HELMERT M., BONET B. & KOENIG S. (2007). Domain-independent construction of pattern database heuristics for cost-optimal planning. In *Proceedings of the 22nd National Conference of the American Association for Artificial Intelligence (AAAI-07)*, p. 1007–1012.

HELMERT M., HASLUM P. & HOFFMANN J. (2007). Flexible abstraction heuristics for optimal sequential planning. In *Proceedings of the 17th International Conference on Automated Planning and Scheduling (ICAPS-07)*, p. 176–183.

KATTENBELT M., KWIATKOWSKA M., NORMAN G. & PARKER D. (2010). A game-based abstraction-refinement framework for Markov Decision Processes. *Form*, **36**(3), 246–280.

KATZ M., HOFFMANN J. & HELMERT M. (2012). How to relax a bisimulation? In *Proceedings of the 22nd International Conference on Automated Planning and Scheduling (ICAPS 2012)*.

NISSIM R., HOFFMANN J. & HELMERT M. (2011). Computing perfect heuristics in polynomial time: On bisimulation and merge-and-shrink abstraction in optimal planning. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, p. 1983–1990.

ORTNER R. (2011). Adaptive aggregation for reinforcement learning in average reward Markov Decision Processes. To appear in Annals of Operations Research.

SANDHOLM T. & SINGH S. (2012). Lossy stochastic game abstraction with bounds. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, p. 880–897: ACM.

WAUGH K., SCHNIZLEIN D., BOWLING M. & SZAFRON D. (2009). Abstraction pathologies in extensive games. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, p. 781–788.

[4] If we aggregate the states 4 and 5 together, in the abstract representation α, we get a bisimulation: all the states behave similarly in regards to each block of the partition $\{1, 2, 3\}, \{4, 5\}$. We have $R(1) = R(2) = R(3) = 1, p(1, \{1, 2, 3\}) = p(2, \{1, 2, 3\}) = p(3, \{1, 2, 3\}) = 0.7$ and $p(1, \{4, 5\}) = p(2, \{4, 5\}) = p(3, \{4, 5\}) = 0.3$, states 4 and 5 are goal states.