Near Optimal Behavior via Approximate State Abstraction

David Abel[†] DAVID_ABEL@BROWN.EDU
D. Ellis Hershkowitz[†] DAVID_HERSHKOWITZ@BROWN.EDU
Michael L. Littman MICHAEL_LITTMAN@BROWN.EDU
Brown University, 115 Waterman Street, Providence, RI 02906

Abstract

The combinatorial explosion that plagues planning and reinforcement learning (RL) algorithms can be moderated using state abstraction. Prohibitively large task representations can be condensed such that essential information is preserved, and consequently, solutions are tractably computable. However, exact abstractions, which treat only fully-identical situations as equivalent, fail to present opportunities for abstraction in environments where no two situations are exactly alike. In this work, we investigate approximate state abstractions, which treat nearly-identical situations as equivalent. We present theoretical guarantees of the quality of behaviors derived from four types of approximate abstractions. Additionally, we empirically demonstrate that approximate abstractions lead to reduction in task complexity and bounded loss of optimality of behavior in a variety of environments.

1. Introduction

Abstraction plays a fundamental role in learning. Through abstraction, intelligent agents may reason about only the salient features of their environment while ignoring what is irrelevant. Consequently, agents are able to solve considerably more complex problems than they would be able to without the use of abstraction. However, *exact abstractions*, which treat only fully-identical situations as equivalent, require complete knowledge that is computationally intractable to obtain. Furthermore, often no two situations are identical, so exact abstractions are often ineffective. To overcome these issues, we investigate *approximate abstractions* that enable agents to treat sufficiently similar situations as identical. This work characterizes the impact of equating "sufficiently similar" states in the context of planning and RL in Markov Decision Processes (MDPs). The remainder of our introduction contextualizes these intuitions in MDPs.

Solving for optimal behavior in MDPs in a planning setting is known to be P-Complete in the size of the state space (Papadimitriou & Tsitsiklis, 1987; Littman et al., 1995). Similarly, many RL algorithms for solving MDPs are known to require a number of samples polynomial in the size of the state space (Strehl et al., 2009). Although polynomial runtime or sample complexity may seem like a reasonable constraint, the size of the state space of an MDP grows superpolynomially with the number of variables that characterize the domain - a result of Bellman's curse of dimensionality. Thus, solutions polynomial in state space size are often ineffective for sufficiently complex tasks. For instance, a robot involved in a pick-and-place task might be able to employ planning algorithms to solve for how to manipulate some objects into a desired configuration in time polynomial in the number of states, but the number of states it must consider grows exponentially with the number of objects with which it is working (Abel et al., 2015).

Thus, a key research agenda for planning and RL is leveraging abstraction to reduce large state spaces (Andre & Russell, 2002; Jong & Stone, 2005; Dietterich, 2000; Bean et al., 2011). This agenda has given rise to methods that reduce *ground* MDPs with large state spaces to *abstract* MDPs with smaller state spaces by aggregating states according to some notion of equality or similarity. In the context of MDPs, we understand exact abstractions as those that aggregate states with equal values of particular quantities, for example, optimal Q-values. Existing work has characterized how exact abstractions can fully maintain optimality in MDPs (Li et al., 2006; Dean & Givan, 1997).

The thesis of this work is that performing approximate abstraction in MDPs by relaxing the state-aggregation criteria from equality to similarity achieves polynomially bounded error in the resulting behavior while offering three benefits. First, approximate abstractions employ the sort of knowl-

Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[†] The first two authors contributed equally.

edge that we expect a planning or learning algorithm to compute without fully solving the MDP. In contrast, exact abstractions often require solving for optimal behavior, thereby defeating the purpose of abstraction. Second, because of their relaxed criteria, approximate abstractions can achieve greater degrees of compression than exact abstractions. This difference is particularly important in environments where no two states are identical. Third, because the state-aggregation criteria are relaxed to near equality, approximate abstractions are able to tune the aggressiveness of abstraction by adjusting what they consider sufficiently similar states.

We support this thesis by describing four different types of approximate abstraction functions that preserve near-optimal behavior by aggregating states on different criteria: $\widetilde{\phi}_{Q^*,\varepsilon}$, on similar optimal Q-values, $\widetilde{\phi}_{\text{model},\varepsilon}$, on similarity of rewards and transitions, $\widetilde{\phi}_{\text{bolt},\varepsilon}$, on similarity of a Boltzmann distribution over optimal Q-values, and $\widetilde{\phi}_{\text{mult},\varepsilon}$, on similarity of a multinomial distribution over optimal Q-values. Furthermore, we empirically demonstrate the relationship between the degree of compression and error incurred on a variety of MDPs.

2. MDPs and Sequential Decision Making

An MDP is a problem representation for sequential decision making agents, represented by a five-tuple: $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$. Here, $\mathcal{S}$ is a finite state space; $\mathcal{A}$ is a finite set of actions available to the agent; $\mathcal{T}$ denotes $\mathcal{T}(s, a, s')$, the probability of an agent transitioning to state $s' \in \mathcal{S}$ after applying action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$; $\mathcal{R}(s, a)$ denotes the reward received by the agent for executing action a in state s; $\gamma \in [0, 1]$ is a discount factor that determines how much the agent prefers future rewards over immediate rewards. We assume without loss of generality that the range of all reward functions is normalized to $[0, 1]$. The solution to an MDP is called a policy, denoted $\pi : \mathcal{S} \mapsto \mathcal{A}$.

The objective of an agent is to solve for the policy that maximizes its expected discounted reward from any state, denoted π^*. We denote the expected discounted reward for following policy π from state s as the value of the state under that policy, $V^\pi(s)$. We similarly denote the expected discounted reward for taking action $a \in \mathcal{A}$ and then following policy π from state s forever after as $Q^\pi(s, a)$, defined by the Bellman Equation as:

$$Q^\pi(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s'} \mathcal{T}(s, a, s') Q^\pi(s', \pi(s')). \quad (1)$$

We let RMAX denote the maximum reward (which is 1), and QMAX denote the maximum Q value, which is $\frac{\text{RMAX}}{1-\gamma}$.

The value function defined under a given policy, denoted $V^\pi(s)$, is defined as:

$$V^\pi(s) = Q^\pi(s, \pi(s)). \quad (2)$$

Lastly, we denote the value and Q functions under the optimal policy as V^* or V^{π^*} and Q^* or Q^{π^*}, respectively. For further background, see Kaelbling et al. (1996).

3. Related Work

Several other projects have addressed similar topics.

3.1. Approximate State Abstraction

Dean et al. (1997) leverage the notion of *bisimulation* to investigate partitioning an MDP's state space into clusters of states whose transition model and reward function are within ε of each other. They develop an algorithm called Interval Value Iteration (IVI) that converges to the correct bounds on a family of abstract MDPs called Bounded MDPs.

Several approaches build on Dean et al. (1997). Ferns et al. (2004; 2006) investigated state similarity metrics for MDPs; they bounded the value difference of ground states and abstract states for several bisimulation metrics that induce an abstract MDP. This differs from our work which develops a theory of abstraction that bounds the suboptimality of applying the optimal policy of an abstract MDP to its ground MDP, covering four types of state abstraction, one of which closely parallels bisimulation. Even-Dar & Mansour (2003) analyzed different distance metrics used in identifying state space partitions subject to ε-similarity. Ortner (2013) developed an algorithm for learning partitions in an online setting by taking advantage of the confidence bounds for $\mathcal{T}$ and $\mathcal{R}$ provided by UCRL (Auer et al., 2009).

3.2. Specific Abstraction Algorithms

Many previous works have targeted the creation of algorithms that enable state abstraction for MDPs. Andre & Russell (2002) investigated a method for state abstraction in hierarchical reinforcement learning leveraging a programming language called ALISP that promotes the notion of *safe* state abstraction. Agents programmed using ALISP can ignore irrelevant parts of the state, achieving abstractions that maintain optimality. Dietterich (2000) developed MAXQ, a framework for composing tasks into an abstracted hierarchy where state aggregation can be applied. Jong & Stone (2005) introduced a method called *policy-irrelevance* in which agents identify (online) which state variables may be safely abstracted away in a factored-state MDP. For a more complete survey of algorithms that leverage state abstraction in past reinforcement-learning papers, see Li et al. (2006).

3.3. Exact Abstraction Framework

Li et al. (2006) developed a framework for exact state abstraction in MDPs. In particular, the authors defined five types of state-aggregation functions, inspired by existing methods for state aggregation in MDPs. We generalize two of these five types, ϕ_{Q^*} and ϕ_{model}, to the approximate abstraction case. Our generalizations are equivalent to theirs when exact criteria are used (i.e. $\varepsilon = 0$). Additionally, when exact criteria are used our bounds indicate that no value is lost, which is one of core results of Li et al. (2006).

4. Abstraction Notation

We build upon the notation used by Li et al. (2006), who introduced a unifying theoretical framework for state abstraction in MDPs.

Definition 1 *($\boldsymbol{M_G}$, $\boldsymbol{M_A}$): We understand an abstraction as a mapping from the state space of a ground MDP, M_G, to that of an abstract MDP, M_A, using a state-aggregation scheme. Consequently, this mapping induces an abstract MDP. Let $M_G = \langle \mathcal{S}_G, \mathcal{A}, \mathcal{T}_G, \mathcal{R}_G, \gamma \rangle$ and $M_A = \langle \mathcal{S}_A, \mathcal{A}, \mathcal{T}_A, \mathcal{R}_A, \gamma \rangle$.*

Definition 2 *($\boldsymbol{\mathcal{S}_A}$, $\boldsymbol{\phi}$): The states in the abstract MDP are constructed by applying a state-aggregation function, ϕ, to the states in the ground MDP, $\mathcal{S}_A$. More specifically, ϕ maps a state in the ground MDP to a state in the abstract MDP:*

$$\mathcal{S}_A = \{\phi(s) \mid s \in \mathcal{S}_G\}. \tag{3}$$

Definition 3 *($\boldsymbol{G}$): Given a ϕ, each ground state has associated with it the ground states with which it is aggregated. Similarly, each abstract state has its constituent ground states. We let G be the function that retrieves these states:*

$$G(s) = \begin{cases} \{g \in \mathcal{S}_G \mid \phi(g) = \phi(s)\}, & \textit{if } s \in \mathcal{S}_G, \\ \{g \in \mathcal{S}_G \mid \phi(g) = s\}, & \textit{if } s \in \mathcal{S}_A. \end{cases} \tag{4}$$

The abstract reward function and abstract transition dynamics for each abstract state are a weighted combination of the rewards and transitions for each ground state in the abstract state.

Definition 4 *($\boldsymbol{\omega(s)}$): We refer to the weight associated with a ground state, $s \in \mathcal{S}_G$ by $\omega(s)$. The only restriction placed on the weighting scheme is that it induces a probability distribution on the ground states of each abstract state:*

$$\forall s \in \mathcal{S}_G \left(\sum_{s \in G(s)} \omega(s) \right) = 1 \quad AND \quad \omega(s) \in [0,1]. \tag{5}$$

Definition 5 *($\boldsymbol{\mathcal{R}_A}$): The abstract reward function $\mathcal{R}_A : \mathcal{S}_A \times \mathcal{A} \mapsto [0,1]$ is a weighted sum of the rewards of each of the ground states that map to the same abstract state:*

$$\mathcal{R}_A(s,a) = \sum_{g \in G(s)} \mathcal{R}_G(g,a)\omega(g). \tag{6}$$

Definition 6 *($\boldsymbol{\mathcal{T}_A}$): The abstract transition function $\mathcal{T}_A : \mathcal{S}_A \times \mathcal{A} \times \mathcal{S}_A \mapsto [0,1]$ is a weighted sum of the transitions of each of the ground states that map to the same abstract state:*

$$\mathcal{T}_A(s,a,s') = \sum_{g \in G(s)} \sum_{g' \in G(s')} \mathcal{T}_G(g,a,g')\omega(g). \tag{7}$$

5. Approximate State Abstraction

Here, we introduce our formal analysis of approximate state abstraction, including results bounding the error associated with these abstraction methods. In particular, we demonstrate that abstractions based on approximate Q^* similarity (5.1), approximate model similarity (5.2), and approximate similarity between distributions over Q^*, for both Boltzmann (5.3) and multinomial (5.4) distributions induce abstract MDPs for which the optimal policy has bounded error in the ground MDP.

We first introduce some additional notation.

Definition 7 *($\boldsymbol{\pi_A^*}$, $\boldsymbol{\pi_G^*}$): We let $\pi_A^* : \mathcal{S}_A \to \mathcal{A}$ and $\pi_G^* : \mathcal{S}_G \to \mathcal{A}$ stand for the optimal policies in the abstract and ground MDPs, respectively.*

We are interested in how the optimal policy in the abstract MDP performs in the ground MDP. As such, we formally define the policy in the ground MDP derived from optimal behavior in the abstract MDP:

Definition 8 *($\boldsymbol{\pi_{GA}}$): Given a state $s \in \mathcal{S}_G$ and a state aggregation function, ϕ,*

$$\pi_{GA}(s) = \pi_A^*(\phi(s)). \tag{8}$$

We now define types of abstraction based on functions of state–action pairs.

Definition 9 *($\boldsymbol{\widetilde{\phi}_{f,\varepsilon}}$): Given a function $f : \mathcal{S}_G \times \mathcal{A} \to \mathbb{R}$ and a fixed non-negative $\varepsilon \in \mathbb{R}$, we define $\widetilde{\phi}_{f,\varepsilon}$ as a type of approximate state-aggregation function that satisfies the following for any two ground states s_1 and s_2:*

$$\widetilde{\phi}_{f,\varepsilon}(s_1) = \widetilde{\phi}_{f,\varepsilon}(s_2) \to \forall_a \, |f(s_1,a) - f(s_2,a)| \leq \varepsilon. \tag{9}$$

That is, when $\widetilde{\phi}_{f,\varepsilon}$ aggregates states, all aggregated states have values of f within ε of each other for all actions.

Definition 10 *($\boldsymbol{Q_G}$, $\boldsymbol{V_G}$): Let $Q_G = Q^{\pi_G^*} : \mathcal{S}_G \times \mathcal{A} \to \mathbb{R}$ and $V_G = V^{\pi_G^*} : \mathcal{S}_G \to \mathbb{R}$ denote the optimal Q and optimal value functions in the ground MDP.*

Definition 11 *($\boldsymbol{Q_A}$, $\boldsymbol{V_A}$): Let $Q_A = Q^{\pi_A^*} : \mathcal{S}_A \times \mathcal{A} \to \mathbb{R}$ and $V_A = V^{\pi_A^*} : \mathcal{S}_A \to \mathbb{R}$ stand for the optimal Q and optimal value functions in the abstract MDP.*

We now introduce our main result.

Theorem 1. *There exist at least four types of approximate state-aggregation functions, $\widetilde{\phi}_{Q^*,\varepsilon}$, $\widetilde{\phi}_{model,\varepsilon}$, $\widetilde{\phi}_{bolt,\varepsilon}$ and $\widetilde{\phi}_{mult,\varepsilon}$, for which the optimal policy in the abstract MDP, applied to the ground MDP, has suboptimality bounded polynomially in ε:*

$$\forall_{s \in \mathcal{S}_G} V_G^{\pi_G^*}(s) - V_G^{\pi_{GA}}(s) \leq poly\left(\varepsilon\right). \tag{10}$$

We prove this theorem in the following sections by proving polynomial bounds on the error of each individual approximate state-aggregation function type.

5.1. Optimal Q Function: $\widetilde{\phi}_{Q^*,\varepsilon}$

We consider an approximate version of Li et al. (2006)'s ϕ_{Q^*}. In our abstraction, states are aggregated together when their optimal Q-values are within ε.

Definition 12 *($\boldsymbol{\widetilde{\phi}_{Q^*,\varepsilon}}$): An approximate Q function abstraction has the same form as Equation 9:*

$$\widetilde{\phi}_{Q^*,\varepsilon}(s_1) = \widetilde{\phi}_{Q^*,\varepsilon}(s_2) \to \forall_a \left|Q_G(s_1, a) - Q_G(s_2, a)\right| \leq \varepsilon. \tag{11}$$

Lemma 1. *When a $\widetilde{\phi}_{Q^*,\varepsilon}$ type abstraction is used to create the abstract MDP:*

$$\forall_{s \in \mathcal{S}_G} V_G^{\pi_G^*}(s) - V_G^{\pi_{GA}}(s) \leq \frac{2\varepsilon}{(1-\gamma)^2}. \tag{12}$$

Proof of Lemma 1: We first demonstrate that Q-values in the abstract MDP are close to Q-values in the ground MDP (Claim 1). We next leverage Claim 1 to demonstrate that the optimal action in the abstract MDP is nearly optimal in the ground MDP (Claim 2). Lastly, we use Claim 2 to conclude Lemma 1 (Claim 3).

Claim 1. *Optimal Q-values in the abstract MDP closely resemble optimal Q-values in the ground MDP:*

$$\forall_{s_G \in \mathcal{S}_G, a} |Q_G(s_G, a) - Q_A(\widetilde{\phi}_{Q^*,\varepsilon}(s_G), a)| \leq \frac{\varepsilon}{1-\gamma}. \tag{13}$$

Consider a non-Markovian decision process of the same form as an MDP, $M_T = \langle \mathcal{S}_T, \mathcal{A}_G, \mathcal{R}_T, \mathcal{T}_T, \gamma \rangle$, parameterized by integer an T, such that for the first T time steps the reward function, transition dynamics and state space are those of the abstract MDP, M_A, and after T time steps the reward function, transition dynamics and state spaces are those of M_G. Thus,

$$\mathcal{S}_T = \begin{cases} \mathcal{S}_G & \text{if } T = 0 \\ \mathcal{S}_A & \text{o/w} \end{cases}$$

$$\mathcal{R}_T(s, a) = \begin{cases} \mathcal{R}_G(s, a) & \text{if } T = 0 \\ \mathcal{R}_A(s, a) & \text{o/w} \end{cases}$$

$$\mathcal{T}_T(s, a, s') = \begin{cases} \mathcal{T}_G(s, a, s') & \text{if } T = 0 \\ \sum\limits_{g \in G(s)} [\mathcal{T}_G(g, a, s')\omega(g)] & \text{if } T = 1 \\ \mathcal{T}_A(s, a, s') & \text{o/w} \end{cases}$$

The Q-value of state s in $\mathcal{S}_T$ for action a is:

$$Q_T(s, a) = \begin{cases} Q_G(s, a) & \text{if } T = 0 \\ \sum\limits_{g \in G(s)} [Q_G(g, a)\omega(g)] & \text{if } T = 1 \\ \mathcal{R}_A(s, a) + \sigma_{T-1}(s, a) & \text{o/w} \end{cases} \tag{14}$$

where:

$$\sigma_{T-1}(s, a) = \gamma \sum_{s_A' \in \mathcal{S}_A} \mathcal{T}_A(s, a, s_A') \max_{a'} Q_{T-1}(s_A', a').$$

We proceed by induction on T to show that:

$$\forall_{T, s_G \in \mathcal{S}_G, a} |Q_T(s_T, a) - Q_G(s_G, a)| \leq \sum_{t=0}^{T-1} \varepsilon\gamma^t, \tag{15}$$

where $s_T = s_G$ if $T = 0$ and $s_T = \widetilde{\phi}_{Q^*,\varepsilon}(s_G)$ otherwise.

Base Case: $T = 0$

When $T = 0$, $Q_T = Q_G$, so this base case trivially follows.

Base Case: $T = 1$
By definition of Q_T, we have that Q_1 is

$$Q_1(s, a) = \sum_{g \in G(s)} \left[Q_G(g, a)\omega(g)\right].$$

Since all co-aggregated states have Q-values within ε of one another and $\omega(g)$ induces a convex combination,

$$Q_1(s_T, a) \leq \varepsilon\gamma^t + \varepsilon + Q_G(s_G, a).$$

$$\therefore |Q_1(s_T, a) - Q_G(s_G, a)| \leq \sum_{t=0}^{1} \varepsilon\gamma^t. \tag{16}$$

Inductive Case: $T > 1$

We assume as our inductive hypothesis that:

$$\forall_{s_G \in \mathcal{S}_G, a} |Q_{T-1}(s_T, a) - Q_G(s_G, a)| \leq \sum_{t=0}^{T-2} \varepsilon\gamma^t.$$

Consider a fixed but arbitrary state, $s_G \in \mathcal{S}_G$, and fixed but arbitrary action a. Since $T > 1$, s_T is $\widetilde{\phi}_{Q^*,\varepsilon}(s_G)$. By definition of $Q_T(s_T, a)$, $\mathcal{R}_A$, $\mathcal{T}_A$:

$$Q_T(s_T, a) = \sum_{g \in G(s_T)} \omega(g) \times$$

$$\left[\mathcal{R}_G(g, a) + \gamma \sum_{g' \in \mathcal{S}_G} \mathcal{T}_G(g, a, g') \max_{a'} Q_{T-1}(g', a')\right].$$

Applying our inductive hypothesis yields:

$$Q_T(s_T, a) \le \sum_{g \in G(s_T)} \omega(g) \times \Bigg[R_G(g, a) + \gamma \sum_{g' \in \mathcal{S}_G} T_G(g, a, g') \max_{a'} (Q_G(g', a') + \sum_{t=0}^{T-2} \varepsilon \gamma^t) \Bigg].$$

Since all aggregated states have Q-values within ε of one another:

$$Q_T(s_T, a) \le \gamma \sum_{t=0}^{T-2} \varepsilon \gamma^t + \varepsilon + Q_G(s_G, a).$$

Since s_G is arbitrary we conclude Equation 15. As $T \to \infty$, $\sum_{t=0}^{T-1} \varepsilon \gamma^t \to \frac{\varepsilon}{1-\gamma}$ by the sum of infinite geometric series and $Q_T \to Q_A$. Thus, Equation 15 yields Claim 1.

Claim 2. *Consider a fixed but arbitrary state, $s_G \in \mathcal{S}_G$ and its corresponding abstract state $s_A = \widetilde{\phi}_{Q^*, \varepsilon}(s_G)$. Let a_G^* stand for the optimal action in s_G, and a_A^* stand for the optimal action in s_A:*

$$a_G^* = \arg\max_a Q_G(s_G, a), \quad a_A^* = \arg\max_a Q_A(s_A, a).$$

The optimal action in the abstract MDP has a Q-value in the ground MDP that is nearly optimal:

$$V_G(s_G) \le Q_G(s_G, a_A^*) + \frac{2\varepsilon}{1-\gamma}. \tag{17}$$

By Claim 1,

$$V_G(s_G) = Q_G(s_G, a_G^*) \le Q_A(s_A, a_G^*) + \frac{\varepsilon}{1-\gamma}. \tag{18}$$

By the definition of a_A^*, we know that

$$Q_A(s_A, a_G^*) + \frac{\varepsilon}{1-\gamma} \le Q_A(s_A, a_A^*) + \frac{\varepsilon}{1-\gamma}. \tag{19}$$

Lastly, again by Claim 1, we know

$$Q_A(s_A, a_A^*) + \frac{\varepsilon}{1-\gamma} \le Q_G(s_g, a_A^*) + \frac{2\varepsilon}{1-\gamma}. \tag{20}$$

Therefore, Equation 17 follows.

Claim 3. *Lemma 1 follows from Claim 2.*

Consider the policy for M_G of following the optimal abstract policy π_A^* for t steps and then following the optimal ground policy π_G^* in M_G:

$$\pi_{A,t}(s) = \begin{cases} \pi_G^*(s) & \text{if } t = 0 \\ \pi_{GA}(s) & \text{if } t > 0 \end{cases} \tag{21}$$

For $t > 0$, the value of this policy for $s_G \in \mathcal{S}_G$ in the ground MDP is:

$$V_G^{\pi_{A,t}}(s_G) = R_G(s, \pi_{A,t}(s_G)) + \gamma \sum_{s_G' \in \mathcal{S}_G} \mathcal{T}_G(s_G, a, s_G') V_G^{\pi_{A,t-1}}(s_G').$$

For $t = 0$, $V_G^{\pi_{A,t}}(s_G)$ is simply $V_G(s_G)$.

We now show by induction on t that

$$\forall_{t, s_G \in \mathcal{S}_g} V_G(s_G) \le V_G^{\pi_{A,t}}(s_G) + \sum_{i=0}^{t} \gamma^i \frac{2\varepsilon}{1-\gamma}. \tag{22}$$

Base case: $t = 0$

By definition, when $t = 0$, $V_G^{\pi_{A,t}} = V_G$, so our bound trivially holds in this case.

Inductive case: $t > 0$

Consider a fixed but arbitrary state $s_G \in \mathcal{S}_G$. We assume for our inductive hypothesis that

$$V_G(s_G) \le V_G^{\pi_{A,t-1}}(s_G) + \sum_{i=0}^{t-1} \gamma^i \frac{2\varepsilon}{1-\gamma}. \tag{23}$$

By definition,

$$V_G^{\pi_{A,t}}(s_G) = R_G(s, \pi_{A,t}(s_G)) + \gamma \sum_{g'} \mathcal{T}_G(s_G, a, s_G') V_G^{\pi_{A,t-1}}(s_G').$$

Applying our inductive hypothesis yields:

$$V_G^{\pi_{A,t}}(s_G) \ge R_G(s_G, \pi_{A,t}(s_G)) + \gamma \sum_{s_G'} \mathcal{T}_G(s_G, \pi_{A,t}(s_G), s_G') \left(V_G(s_G') - \sum_{i=0}^{t-1} \gamma^i \frac{2\varepsilon}{1-\gamma} \right).$$

Therefore,

$$V_G^{\pi_{A,t}}(s_G) \ge -\gamma \sum_{i=0}^{t-1} \gamma^i \frac{2\varepsilon}{1-\gamma} + Q_G(s_G, \pi_{A,t}(s_G)).$$

Applying Claim 2 yields:

$$V_G^{\pi_{A,t}}(s_G) \ge -\gamma \sum_{i=0}^{t-1} \gamma^i \frac{2\varepsilon}{1-\gamma} - \frac{2\varepsilon}{1-\gamma} + V_G(s_G)$$

$$\therefore V_G(s_G) \le V_G^{\pi_{A,t}}(s_G) + \sum_{i=0}^{t} \gamma^i \frac{2\varepsilon}{1-\gamma}.$$

Since s_G was arbitrary, we conclude that our bound holds for all states in $\mathcal{S}_G$ for the inductive case. Thus, from our base case and induction, we conclude that

$$\forall_{t, s_G \in \mathcal{S}_g} V_G^{\pi_G^*}(s_G) \le V_G^{\pi_{A,t}}(s_G) + \sum_{i=0}^{t} \gamma^i \frac{2\varepsilon}{1-\gamma}. \tag{24}$$

Note that as $t \to \infty$, $\sum_{i=0}^{t} \gamma^i \frac{2\varepsilon}{1-\gamma} \to \frac{2\varepsilon}{(1-\gamma)^2}$ by the sum of infinite geometric series and $\pi_{A,t}(s) \to \pi_{GA}$. Thus, we conclude Lemma 1. □

5.2. Model Similarity: $\widetilde{\phi}_{model,\varepsilon}$

Now, consider an approximate version of Li et al. (2006)'s ϕ_{model}, where states are aggregated together when their rewards and transitions are within ε.

Definition 13 *($\widetilde{\phi}_{model,\varepsilon}$): We let $\widetilde{\phi}_{model,\varepsilon}$ define a type of abstraction that, for fixed ε, satisfies:*

$$\begin{gathered}\widetilde{\phi}_{model,\varepsilon}(s_1) = \widetilde{\phi}_{model,\varepsilon}(s_2) \to \\ \forall_a \, |\mathcal{R}_G(s_1,a) - \mathcal{R}_G(s_2,a)| \le \varepsilon \text{ AND} \\ \forall_{s_A \in \mathcal{S}_A} \left| \sum_{s_G{}' \in G(s_A)} [\mathcal{T}_G(s_1,a,s_G{}') - \mathcal{T}_G(s_2,a,s_G{}')] \right| \le \varepsilon.\end{gathered} \tag{25}$$

Lemma 2. *When $\mathcal{S}_A$ is created using a $\widetilde{\phi}_{model,\varepsilon}$ type:*

$$\forall_{s \in \mathcal{S}_G} V_G^{\pi_G^*}(s) - V_G^{\pi_{GA}}(s) \le \frac{2\varepsilon + 2\gamma((|\mathcal{S}_G| - 1)\varepsilon)}{(1-\gamma)^3}. \tag{26}$$

Proof of Lemma 2:

Let B be the maximum Q-value difference between any pair of ground states in the same abstract state for $\widetilde{\phi}_{\text{model},\varepsilon}$:

$$B = \max_{s_A, s_1, s_2, a} |Q_G(s_1,a) - Q_G(s_2,a)|,$$

where $s_A \in \mathcal{S}_A$ and $s_1, s_2 \in G(s_A)$. Since difference of rewards is bounded by ε:

$$B \le \varepsilon + \gamma \sum_{s_A \in \mathcal{S}_A} \sum_{s_G{}' \in G(s_A)} \Big[(T_G(s_1,a,s_G{}') - T_G(s_2,a,s_G{}')) \max_{a'} Q_G(s_G{}', a') \Big].$$

By similarity of transitions under $\widetilde{\phi}_{\text{model},\varepsilon}$:

$$B \le \varepsilon + \gamma \text{QM}\textsc{ax} \sum_{s_A \in \mathcal{S}_A} \varepsilon \le \varepsilon + \gamma |\mathcal{S}_G| \varepsilon \text{QM}\textsc{ax}.$$

Since $\text{QM}\textsc{ax} = \frac{\text{RM}\textsc{ax}}{1-\gamma}$, and we defined $\text{RM}\textsc{ax} = 1$:

$$B \le \frac{\varepsilon + \gamma(|\mathcal{S}_G| - 1)\varepsilon}{1-\gamma}.$$

Since the Q-values of ground states grouped under $\widetilde{\phi}_{\text{model},\varepsilon}$ are strictly less than B, we can understand $\widetilde{\phi}_{\text{model},\varepsilon}$ as a type of $\widetilde{\phi}_{Q^*,B}$. Applying Lemma 1 yields Lemma 2. □

5.3. Boltzmann over Optimal Q: $\widetilde{\phi}_{\text{bolt},\varepsilon}$

Here, we introduce $\widetilde{\phi}_{\text{bolt},\varepsilon}$, which aggregates states with similar Boltzmann distributions on Q-values. This family of abstractions is appealing as Boltzmman distributions balance exploration and exploitation (Sutton & Barto, 1998). We find this type particularly interesting for abstraction purposes as, unlike $\widetilde{\phi}_{Q^*,\varepsilon}$, it allows for aggregation when Q-value ratios are similar but their magnitudes are different.

Definition 14 *($\widetilde{\phi}_{bolt,\varepsilon}$): We let $\widetilde{\phi}_{bolt,\varepsilon}$ define a type of abstractions that, for fixed ε, satisfies:*

$$\begin{gathered}\widetilde{\phi}_{bolt,\varepsilon}(s_1) = \widetilde{\phi}_{bolt,\varepsilon}(s_2) \to \\ \forall_a \left| \frac{e^{Q_G(s_1,a)}}{\sum_b e^{Q_G(s_1,b)}} - \frac{e^{Q_G(s_2,a)}}{\sum_b e^{Q_G(s_2,b)}} \right| \le \varepsilon.\end{gathered} \tag{27}$$

We also assume that the difference in normalizing terms is bounded by some non-negative constant, $k \in \mathbb{R}$, of ε:

$$\left| \sum_b e^{Q_G(s_1,b)} - \sum_b e^{Q_G(s_2,b)} \right| \le k \times \varepsilon. \tag{28}$$

Lemma 3. *When S_A is created using a function of the $\widetilde{\phi}_{bolt,\varepsilon}$ type, for some non-negative constant $k \in \mathbb{R}$:*

$$\forall_{s \in \mathcal{S}_G} V_G^{\pi_G^*}(s) - V_G^{\pi_{GA}}(s) \le \frac{2\varepsilon\left(\frac{|\mathcal{A}|}{1-\gamma} + k\varepsilon + k\right)}{(1-\gamma)^2}. \tag{29}$$

We use the approximation for e^x, with δ error:

$$e^x = 1 + x + \delta \approx 1 + x. \tag{30}$$

We let δ_1 denote the error in approximating $e^{Q_G(s_1,a)}$ and δ_2 denote the error in approximating $e^{Q_G(s_2,a)}$.

Proof Sketch of Lemma 3:

By the approximation in Equation 30 and the assumption in Equation 28:

$$\left| \frac{1 + Q_G(s_1,a) + \delta_1}{\sum_j e^{Q_G(s_1,a_j)}} - \frac{1 + Q_G(s_2,a) + \delta_2}{\sum_j e^{Q_G(s_1,a_j)} \pm k\varepsilon} \right| \le \varepsilon \tag{31}$$

Either $\sum_j e^{Q_G(s_1,a_j)} \pm k\varepsilon$ is $\sum_j e^{Q_G(s_1,a_j)} + k\varepsilon$, or $\sum_j e^{Q_G(s_1,a_j)} - k\varepsilon$.

First suppose the former. It follows by algebra that:

$$-\varepsilon \le \frac{1 + Q_G(s_1,a) + \delta_1}{\sum_j e^{Q_G(s_1,a_j)}} - \frac{1 + Q_G(s_2,a) + \delta_2}{\sum_j e^{Q_G(s_1,a_j)} + k\varepsilon} \le \varepsilon \tag{32}$$

$$-\varepsilon\left(k\varepsilon+\sum_j e^{Q_G(s_1,a_j)}\right)-\delta_1+\delta_2\le$$

$$k\varepsilon\left(\frac{1+Q_G(s_1,a)+\delta_1}{\sum_j e^{Q_G(s_1,a_j)}}\right)+Q_G(s_1,a)-Q_G(s_2,a)\le$$

$$\varepsilon\left(k\varepsilon+\sum_j e^{Q_G(s_1,a_j)}\right)-\delta_1+\delta_2$$

A similar equation follows if we suppose the latter and combining these equations gives us:

$$|Q_G(s_1,a)-Q_G(s_2,a)|\le\varepsilon\left(\frac{|\mathcal{A}|}{1-\gamma}+k\varepsilon+k\right). \quad (33)$$

Consequently, we can consider $\widetilde{\phi}_{\text{bolt},\varepsilon}$ as a special case of the $\widetilde{\phi}_{Q^*,B}$ type, where $B=\varepsilon\left(\frac{|\mathcal{A}|}{1-\gamma}+k\varepsilon+k\right)$. Lemma 3 then follows from Lemma 1. □

5.4. Multinomial over Optimal Q: $\widetilde{\phi}_{\textbf{mult},\varepsilon}$

We consider approximate abstractions derived from a multinomial distribution over Q^* for similar reasons to the Boltzmann distribution. Additionally, the multinomial distribution is appealing for its simplicity.

Definition 15 *($\widetilde{\phi}_{\textbf{mult},\varepsilon}$): We let $\widetilde{\phi}_{\text{mult},\varepsilon}$ define a type of abstraction that, for fixed ε, satisfies*

$$\widetilde{\phi}_{\text{mult},\varepsilon}(s_1)=\widetilde{\phi}_{\text{mult},\varepsilon}(s_2)\rightarrow$$
$$\forall_a\left|\frac{Q_G(s_1,a)}{\sum_b Q_G(s_1,b)}-\frac{Q_G(s_1,a)}{\sum_b Q_G(s_1,b)}\right|\le\varepsilon. \quad (34)$$

We also assume that the difference in normalizing terms is bounded by some non-negative constant, $k\in\mathbb{R}$, of ε:

$$\left|\sum_i Q_G(s_1,a_i)-\sum_j Q_G(s_2,a_j)\right|\le k\times\varepsilon. \quad (35)$$

Lemma 4. *When S_A is created using a function of the $\widetilde{\phi}_{\text{mult},\varepsilon}$ type, for some non-negative constant $k\in\mathbb{R}$:*

$$\forall_{s\in S_M}V_G^{\pi_G^*}(s)-V_G^{\pi_{GA}}(s)\le\frac{\frac{2\varepsilon|\mathcal{A}|}{1-\gamma}+k\varepsilon^2+k}{(1-\gamma)^2}. \quad (36)$$

Proof Sketch of Lemma 4 The proof follows an identical strategy to that of Lemma 3, but without $e^x\approx 1+x$. □

6. Example Domains

We apply approximate abstraction to four example domains—NChain, Taxi, Minefield and Random. These domains were selected for their diversity—NChain is relatively simple, Taxi is goal-based and hierarchical in nature, Minefield is stochastic, and Random MDP has many near-optimal policies.

Our code base[1] provides implementations for abstracting arbitrary MDPs as well as visualizing and evaluating the resulting abstract MDPs. We use the graph-visualization library GraphStream (Pigné et al., 2008) and the planning and RL library, BURLAP[2]. For all experiments, we set γ to 0.95.

NChain is a simple MDP investigated in the Bayesian RL literature due to the interesting exploration problem it poses (Dearden et al., 1998). In our implementation, we set $N=10$, normalized rewards between 0 and 1, and used a slip probability of 0.2.

Taxi has long been studied by the hierarchical RL literature (Dietterich, 2000). The agent, operating in a Grid World style domain (Russell & Norvig, 1995), may move left, right, up, and down, as well as pick up a passenger and drop off a passenger. The goal is achieved when the agent has taken all passengers to their destinations.

Minefield is a test problem we are introducing that uses the Grid World dynamics of Russell & Norvig (1995) with slip probability of x. The reward function is such that moving up in the top row of the grid receives 1.0 reward; all other transitions receive 0.2 reward, except for transitions to a random set of κ states (which may include the top row) that receive 0 reward. (These are the states with "mines" in them.) We set $N=10, M=4, \varepsilon=0.5, \kappa=5, x=0.01$. In the Random MDP domain we consider, there are 100 states and 3 actions. For each state, each action transitions to one of two randomly selected states with probability 0.5.

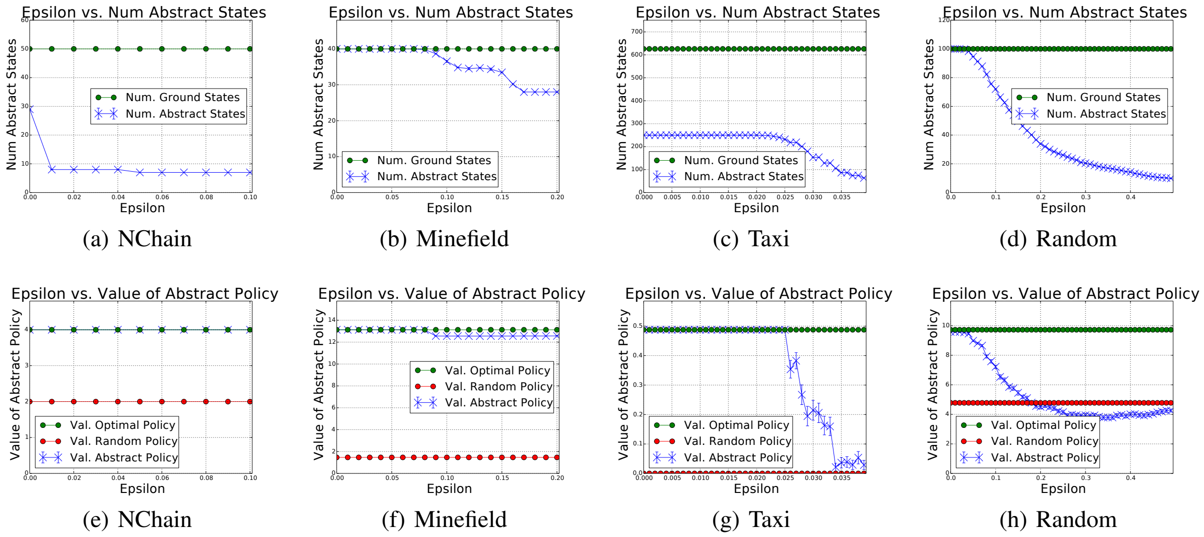
7. Empirical Results

We ran experiments on the $\widetilde{\phi}_{Q^*,\varepsilon}$ type aggregation functions. We provide results for only $\widetilde{\phi}_{Q^*,\varepsilon}$ because, as our proofs in Section 5 demonstrate, the other three functions are reducible to particular $\widetilde{\phi}_{Q^*,\varepsilon}$ functions. For the purpose of illustrating what kinds of approximations are possible we built each abstraction by first solving the MDP, then greedily aggregating ground states into abstract states that satisfied the $\widetilde{\phi}_{Q^*,\varepsilon}$ criteria. Since this approach represents an order-dependent approximation to the maximum amount of abstraction possible, we randomized the order in which states were considered across trials. Every ground state is equally weighted in its abstract state.

For each domain, we report two quantities as a function of epsilon with 95% confidence bars. First, we compare the number of states in the abstract MDP for different values

[1] github.com/david-abel/state_abstraction

[2] http://burlap.cs.brown.edu/



(a) NChain

(b) Minefield

(c) Taxi

(d) Random

(e) NChain

(f) Minefield

(g) Taxi

(h) Random

Figure 1. ε vs. Num States and ε vs. Abstract Policy Value

of ε, shown in the top row of Figure 1. The smaller the number of abstract states, the smaller the state space of the MDP that the agent must plan over. Second, we report the value under the abstract policy of the initial ground state, also shown in the bottom row of Figure 1. In the Taxi and Random domains, 200 trials were run for each data point, whereas 20 trials were sufficient in Minefield and NChain.

Our empirical results corroborate our thesis—approximate state abstractions can decrease state space size while retaining bounded error. In both NChain and Minefield, we observe that, as ε increases from 0, the number of states that must be planned over is reduced, and optimal behavior is either fully maintained (NChain) or very nearly maintained (Minefield). Similarly for Taxi, when ε is between .02 and .025, we observe a reduction in the number of states in the abstract MDP while value is fully maintained. After .025, increased reduction in state space size comes at a cost of value. Lastly, as ε is increased in the Random domain, there is a smooth reduction in the number of abstract states with a corresponding cost in the value of the derived policy. When $\varepsilon = 0$, there is no reduction in state space size whatsoever (the ground MDP has 100 states), because no two states have identical optimal Q-values.

Our experimental results also highlight a noteworthy characteristic of approximate state abstraction in goal-based MDPs. Taxi exhibits relative stability in state space size and behavior for ε up to .02, at which point both fall off dramatically. We attribute the sudden fall off of these quantities to the goal-based nature of the domain; once information critical for achieving optimal behavior is lost in the state aggregation, solving the goal—and so acquiring any reward—is impossible. Conversely, in the Random domain, a great deal of near optimal policies are available to the agent. Thus, even as the information for optimal behavior is lost, there are many near optimal policies available to the agent that remain available.

8. Conclusion

Approximate abstraction in MDPs offers considerable advantages over exact abstraction. In this work, we proved bounds for the value lost when behaving according to the optimal policy of the abstract MDP. We also empirically demonstrate that approximate abstractions can reduce state space size with minor loss in the quality of the behavior.

There are many directions for future work. First, we are interested in extending the approach of Ortner (2013) by learning the approximate abstraction functions introduced in this paper online in the planning or RL setting. While our work presents several sufficient conditions for achieving bounded error of learned behavior with approximate abstractions, we hope to investigate what conditions are strictly necessary for an approximate abstraction to achieve bounded error. In the future, we are interested in characterizing the relationship between temporal abstractions, such as options (Sutton et al., 1999), and approximate abstractions. Lastly, we are interested in understanding the relationship between various approximate abstractions and the information theoretical limitations on the degree of abstraction achievable in MDPs.

References

Abel, David, Hershkowitz, David Ellis, Barth-Maron, Gabriel, Brawner, Stephen, O'Farrell, Kevin, MacGlashan, James, and Tellex, Stefanie. Goal-based action priors. In *ICAPS*, pp. 306–314, 2015.

Andre, David and Russell, Stuart J. State abstraction for programmable reinforcement learning agents. In *AAAI/IAAI*, pp. 119–125, 2002.

Auer, Peter, Jaksch, Thomas, and Ortner, Ronald. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 89–96, 2009.

Bean, James C, Birge, John R, and Smith, Robert L. Dynamic programming aggregation. *Operations Research*, 35(2):215–220, 2011.

Dean, Thomas and Givan, Robert. Model minimization in markov decision processes. In *AAAI/IAAI*, pp. 106–111, 1997.

Dean, Thomas, Givan, Robert, and Leach, Sonia. Model reduction techniques for computing approximately optimal solutions for markov decision processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 124–131. Morgan Kaufmann Publishers Inc., 1997.

Dearden, Richard, Friedman, Nir, and Russell, Stuart. Bayesian Q-learning. In *AAAI/IAAI*, pp. 761–768, 1998.

Dietterich, Thomas G. Hierarchical reinforcement learning with the maxq value function decomposition. *J. Artif. Intell. Res.(JAIR)*, 13:227–303, 2000.

Even-Dar, Eyal and Mansour, Yishay. Approximate equivalence of Markov decision processes. In *Learning Theory and Kernel Machines*, pp. 581–594. Springer, 2003.

Ferns, Norm, Panangaden, Prakash, and Precup, Doina. Metrics for finite markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 162–169. AUAI Press, 2004.

Ferns, Norman, Castro, Pablo Samuel, Precup, Doina, and Panangaden, Prakash. Methods for computing state similarity in markov decision processes. *Proceedings of the 22nd conference on Uncertainty in artificial intelligence*, 2006.

Jong, Nicholas K and Stone, Peter. State abstraction discovery from irrelevant state variables. In *IJCAI*, pp. 752–757, 2005.

Kaelbling, Leslie Pack, Littman, Michael L, and Moore, Andrew W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, pp. 237–285, 1996.

Li, Lihong, Walsh, Thomas J, and Littman, Michael L. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

Littman, Michael L, Dean, Thomas L, and Kaelbling, Leslie Pack. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 394–402. Morgan Kaufmann Publishers Inc., 1995.

Ortner, Ronald. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. *Annals of Operations Research*, 208(1):321–336, 2013.

Papadimitriou, Christos H and Tsitsiklis, John N. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

Pigné, Yoann, Dutot, Antoine, Guinand, Frédéric, and Olivier, Damien. Graphstream: A tool for bridging the gap between complex systems and dynamic graphs. *CoRR*, abs/0803.2093, 2008.

Russell, Stuart and Norvig, Peter. *Artificial Intelligence A Modern Approach*. Prentice-Hall, Englewood Cliffs, 1995.

Strehl, Alexander L, Li, Lihong, and Littman, Michael L. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Sutton, Richard S and Barto, Andrew G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Sutton, Richard S, Precup, Doina, and Singh, Satinder. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999.