Performance Guarantees for Homomorphisms beyond Markov Decision Processes

Sultan Javed Majeed,[1] **Marcus Hutter**[2]
[1,2]Research School of Computer Science, Australian National University, Australia
[1]http://www.sultan.pk, [2]http://www.hutter1.net

Abstract

Most *real-world* problems have huge state and/or action spaces. Therefore, a naive application of existing *tabular* solution methods is not tractable on such problems. Nonetheless, these solution methods are quite useful if an agent has access to a relatively small state-action space homomorphism of the true environment and near-optimal performance is guaranteed by the map. A plethora of research is focused on the case when the homomorphism is a Markovian representation of the underlying process. However, we show that near-optimal performance is sometimes guaranteed even if the homomorphism is non-Markovian.

Introduction

The task of learning a near-optimal behavior from a sequence of experiences can naturally be formulated as a Reinforcement Learning (RL) problem (Sutton and Barto 2018). In a typical RL framework, an agent interacts with an environment by taking an action and receiving a feedback.

It is typically assumed that the agent is facing a small state-action space[1] Markov Decision Process (MDP) so the agent can advise a stationary policy as a function of state (Puterman 2014). Unfortunately, the number of state-action pairs in most of real-world problems is prohibitively large, e.g. driving a car, playing Go, personal assistance, controlling a plant with real-valued inputs, and so forth. The agent can neither simply visit each state-action pair nor can it keep record of these visits to learn a near-optimal behavior. This explosion of state-action space is known as the *curse of dimensionality* (Sutton and Barto 2018). Therefore, it is essential for the agent to generalize over its experiences in such a huge state-action space problem.

The curse of dimensionality is not a mere artifact of limited experience and computation constraints if the problem has an infinite state-action space. In a typical General Reinforcement Learning[2] (GRL) framework the agent is faced with an environment without any known structure. The GRL setup is arguably the most general setup: it can represent MDP, $k-$MDP, partially observed MDP (POMDP) and other typical environment models (Hutter 2016; Leike 2016). But this generality comes at the cost of an infinite state-action space: every agent-environment interaction generates a *unique* history. Hence, there is no other option but to consider every history as a unique state of the environment. Therefore, GRL suffers, inevitably, from the curse of dimensionality.

A *homomorphism* framework originated by Whitt (1978) is a well-studied solution to handle the state-action space curse of dimensionality. In the homomorphism framework a problem of a large state-action space is *solved* by using an *abstract* problem with a relatively small state-action space. The (near-)optimal policy of the abstract problem is a *solution* if it is also a (near-)optimal policy in the true environment.

It is important to highlight that homomorphism is not the only technique for abstracting actions. The *options* framework is a competing method for temporal action abstractions (Sutton, Precup, and Singh 1999). In the option/macro-action framework, the original action space is augmented with long-term/built-in policies (McGovern, Sutton, and Fagg 1997). The agent using an option/marco-action commits to execute a fixed set of actions for a fixed (expected) time duration. This temporal action abstraction framework is arguably more powerful but beyond the scope of this work. Because, to the best of our knowledge, there are no theoretical performance guarantees available for such methods, and most probably such bounds might not exist.

In the homomorphism framework, it is typically assumed that the abstract problem is an MDP (Ravindran and Barto 2003; 2004; Taylor, Precup, and Panangaden 2008). However, the size of the abstract state-action space can be significantly reduced if non-MDP abstractions are possible (Abel, Hershkowitz, and Littman 2016; Li, Walsh, and Littman 2006). Moreover, the reduction of abstract state-action space roughly translates into faster learning and planning[3] (Strehl, Li, and Littman 2009; Lattimore and Hutter 2014).

It has recently been shown that the MDP restriction is not a necessary condition for near-optimal performance guarantees in state-only abstractions of GRL (Hutter 2016). In

[1] We refer a state and action space/pair jointly as a *state-action* space/pair.

[2] We formally define a typical GRL setup in Preliminaries.

[3] Although reduction of state-action space is necessary for faster learning/planning but not sufficient (Littman, Dean, and Kaelbling 1995).

this work, we use similar notation and techniques of Hutter (2016) but investigate and prove optimality bounds for non-MDP state-action homomorphisms in GRL. Since state abstraction is a special case of homomorphism (the action space is not reduced/mapped), our work is a generalization of Extreme State Aggregation (ESA) (Hutter 2016).

The homomorphism framework has been extended beyond MDPs to finite-state POMDPs (Wolfe 2010). As mentioned earlier, GRL has an infinite set of histories and no two histories are alike. We can represent a finite-state POMDP environment as a history-based process by imposing a structure that there is an internal MDP that generates the observations and rewards. The GRL framework, by design, is more powerful and expressive than a finite-state POMDP (Wolfe 2010; Leike 2016). Therefore, our results are more general than finite-state POMDP homomorphisms.

Preliminaries

This section provides the required notation, a typical GRL framework and our homomorphism setup. We consider a simple agent-environment setup (Sutton and Barto 2018). The agent has a finite set of actions $\mathcal{A}$. The environment receives an action from the agent and gives a standard observation from a finite set of observations $\mathcal{O}$ and a real-valued reward from a finite set $\mathcal{R} \subseteq \mathbb{R}$. The agent interacts with the environment in cycles, and in each cycle the agent performs an action and receives an observation and reward from the environment. This agent-environment interaction generates a possibly infinite history from an infinite set of histories $\mathcal{H} := \bigcup_{t=0}^{\infty} (\mathcal{A} \times \mathcal{O} \times \mathcal{R})^t$. Hence, the original state-action space is the history-action space[4], i.e. $\mathcal{H} \times \mathcal{A}$. Similarly, we define an abstract finite state space $\mathcal{S}$ and action space $\mathcal{B}$ to form the abstract state-action space, i.e. $\mathcal{S} \times \mathcal{B}$.

We use a consistent notation throughout the paper unless stated otherwise. We use $\Delta(\cdot)$ to denote a probability distribution over its argument, $\|\cdot\|_1$ expresses the 1-norm, $\tilde{x}$ is a local variable and x' is a different member of the same set. We use a shorthand notation $\forall f(x) = y$ to imply $\forall x, y : f(x) = y$. We often make references to the results presented later in the paper. The reader is not encouraged to follow these *justifying* references during the first reading.

General Reinforcement Learning Framework

This section provides a formal layout of a typical GRL framework and some assumptions we make about the setup. We start our setup by defining two center pieces of any RL setup: the environment and the agent/policy[5]. The environment, also referred as the *original process* P, is defined as a stochastic mapping from a history-action pair to a distribution over the observation-reward pairs, i.e. $P : \mathcal{H} \times \mathcal{A} \to \Delta(\mathcal{O} \times \mathcal{R})$. The *history-based* agent/policy Π is defined to be a function that stochastically maps a history to the actions as $\Pi : \mathcal{H} \to \Delta(\mathcal{A})$.

Assumption 1. (Geometric discounting) *We assume a geometric discounting over the rewards — i.e. the agent discounts its future rewards by a constant discount factor* $\gamma \in [0, 1)$.

The goal of the agent is to maximize the expected discounted sum of rewards which is generally expressed with Bellman equations of (action-)value functions (Sutton and Barto 2018). The agent tries to maximize this value function and strives to reach the most valuable states. We define the action-value function Q^Π for any history $h \in \mathcal{H}$ and action $a \in \mathcal{A}$ as

$$Q^\Pi(h, a) := \sum_{\tilde{o} \in \mathcal{O}, \tilde{r} \in \mathcal{R}} P(\tilde{o}\tilde{r}|ha) \left(\tilde{r} + \gamma V^\Pi(\tilde{h})\right) \quad (1)$$

where $\tilde{h} := ha\tilde{o}\tilde{r}$ is an extended history and the corresponding value function V^Π is defined as

$$V^\Pi(h) := \sum_{\tilde{a} \in \mathcal{A}} Q^\Pi(h, \tilde{a}) \Pi(\tilde{a}|h). \quad (2)$$

The (action-)value functions are maximized if the agent is following an *optimal policy* $\Pi^* \in \arg\max_{\tilde{\Pi}} V^{\tilde{\Pi}}$.

Assumption 2. (Bounded positive reward) *We assume bounded and positive rewards and without loss of generality we assume* $\mathcal{R} :\subseteq [0, 1]$.

It is easy to see that the bounded rewards bound the (action-)value functions between 0 and $1/(1 - \gamma)$.

Homomorphism Setup

We define a homomorphism as a surjective mapping ψ from the original state-action space $\mathcal{H} \times \mathcal{A}$ to the abstract state-action space $\mathcal{S} \times \mathcal{B}$.

For a succinct exposition, we also define a few marginalized mapping functions. These marginalized maps do not have any special significance other than making the notation a bit simpler.

Histories mapped to an sb-pair. For a given abstract action $b \in \mathcal{B}$, we define a marginalized abstract state map as

$$\psi_b^{-1}(s) := \{h \in \mathcal{H} \mid \exists a \in \mathcal{A} : \psi(h, a) = (s, b)\}. \quad (3)$$

Actions mapped to an sb-pair. Similarly, we also define a marginalized abstract action map for any abstract state $s \in \mathcal{S}$ and history $h \in \mathcal{H}$ as

$$\psi_s^{-1}(b) := \{a \in \mathcal{A} \mid \psi(h, a) = (s, b)\}. \quad (4)$$

It is important to note that $\psi_s^{-1}(b)$ is also a function of history. This dependence is always clear from the context, so we suppress it in the notation.

Abstract states mapped by a history. By a slight abuse of notation we overload ψ, and define a history to abstract state marginalized map as

$$\psi(h) := \{s \in \mathcal{S} \mid \exists a \in \mathcal{A}, b \in \mathcal{B} : \psi(h, a) = (s, b)\}. \quad (5)$$

Histories mapped to an abstract state. Finally, an abstract state to history marginalized map is defined as

$$\psi^{-1}(s) := \{h \in \mathcal{H} \mid \exists a \in \mathcal{A}, b \in \mathcal{B} : \psi(h, a) = (s, b)\}. \quad (6)$$

[4] In general, histories are considered as the states of the environment, so we interchangeably call the history-action space the *original* state-action space.

[5] While it can be argued that an agent and a policy are two separate entities, in this work we use them interchangeably.

Assumption 3. ($\psi(h) = s$) *We assume that an abstract state is determined only by the history — i.e. $\psi(h, a) := (s = f(h), b)$, where f is any fixed* surjective *function of history and is independent of actions a and b.*

The above assumption implies that $\psi(h)$ is *singleton*. This is not only a technical necessity but a requirement to make the mapping *causal*, i.e. the current history h corresponds to a unique state s independent of the next action taken by the agent. If we drop this assumption then the current history might resolve to a different state based on the next (future) action taken by the agent.

A homomorphic map ψ lets the agent merge the experiences from P and induces a *history-based* abstract process P_ψ. Formally, for all $\psi(h, a) = (s, b)$ and any next abstract state s', we express P_ψ as

$$P_\psi(s'r|ha) := \sum_{\tilde{o}:\psi(ha\tilde{o}r)=s'} P(\tilde{o}r|ha). \tag{7}$$

The map ψ also induces a *history-based* abstract policy Π_ψ as

$$\Pi_\psi(b|h) := \sum_{\tilde{a}\in\psi_s^{-1}(b)} \Pi(\tilde{a}|h). \tag{8}$$

It is clear from (7) and (8) that the induced abstract process and policy are in general non-Markovian, i.e. both are functions of the history h and not only the abstract state s.

Non-MDP homomorphisms. In this work we consider two types of non-Markovian homomorphisms: **a)** *Q-uniform* homomorphisms, where the state-action pairs are merged if they have close Q-values, i.e. $Q^\Pi(h, a) \approx Q^\Pi(h', a')$ for all $\psi(h, a) = \psi(h', a')$, and **b)** *V-uniform* homomorphisms, when the merged state-action pairs have close values , i.e. $V^\Pi(h) \approx V^\Pi(h')$ for all $\psi(h) = \psi(h')$. A formal treatment of these non-MDP homomorphisms is provided in the main results section. In both Q and V-uniform homomorphisms, P_ψ can be history-dependent, i.e. the abstract process is non-MDP.

Motivation for Non-MDP Homomorphisms

In this section we motivate the importance of non-MDP homomorphisms by an example. We show that a non-MDP homomorphism can cater to a large set of domains and allows more compact representations.

Navigational Grid-world. Let us consider a simplified version of the asymmetric grid-world example by Ravindran and Barto (2004) in Figure 1. In this navigational domain, the goal of an agent Π is to navigate the grid to reach the target cell T. The unreachable cells are grayed-out. The agent receives a large positive reward if it enters the cell T, otherwise a small negative reward is given to the agent at each time-step. The agent is capable of moving in the four directions, i.e. up, down, left and right. This domain has an *almost* similar transition and reward structure across a diagonal axis. We call this an approximate MDP axis and denote it by ≈MDP. This axis of symmetry enables us to create a homomorphism of the domain using approximately half of the original state-space (see Figure 2).

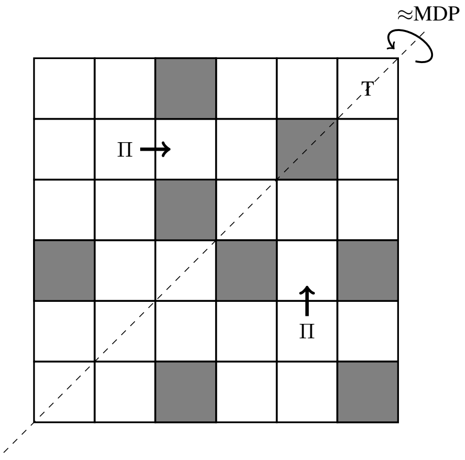


Figure 1: The original navigational grid-world with the axis of approximate symmetry. The gray cells are not reachable. The target cell is at the top right corner. The figure shows two possible positions of the agent and corresponding optimal actions.

This grid-world example has primarily been studied in the context of either exact, approximate or Bounded parameter MDP (BMDP) homomorphisms (Ravindran and Barto 2004): the abstract model *approximately* preserves the one-step dynamics of the original environment. However, as we later prove in this paper (see Theorem 8ii), some non-MDP homomorphisms can also be used to find a near-optimal policy in the original process. We motivate the need of non-MDP homomorphisms, first, by highlighting the fact[6] that in the grid-world domain, the states with similar dynamics have similar optimal action-values. Afterwards, we modify the grid-world domain such that the modified grid-world does not have an approximate MDP symmetry axis, but still has the same approximate optimal action-values symmetry.

We apply Value Iteration (VI) (Bellman 1957) with some fixed but irrelevant parameters on the grid world (see Figure 3). The grid world has the same approximate symmetry axis for the optimal values, denoted by ≈Q-uniform axis. It is easy to see that each merged state in Figure 2 has the same action-values. Hence, the ≈MDP axis is also an ≈Q-uniform axis in the grid-world.

Modified Navigational Grid-world. Now we modify the grid world such that it does not have an ≈MDP axis (Figure 1) but it still has the same ≈Q-uniform axis (Figure 3). The idea is to take a pair of merged states from Figure 1 and change the reward and transition probabilities such that the states no longer have similar one-step dynamics but still have similar action-values. For example, let us consider the cells highlighted with dashed borders in Figure 3 and denote the cell in the bottom half with s_{23}. Let u, d, p_u and p_d denote the actions up and down, and the probabilities to reach the desired cell by taking the corresponding action, respectively. Let r_u and r_d be the expected rewards for each action

[6] This section is an informal motivation, we formally deal with this fact in the main results section (Theorem 6i).

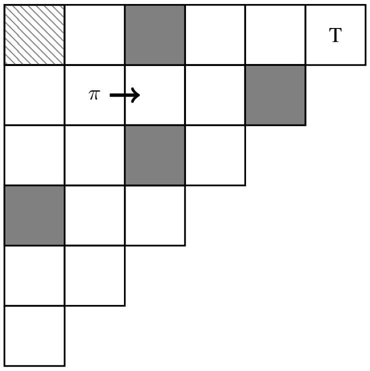
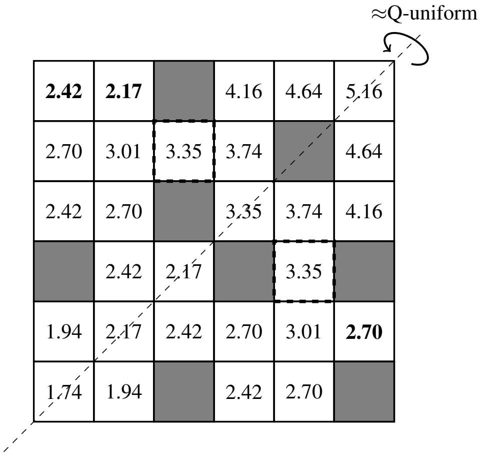


Figure 2: A possible MDP homomorphism by merging the mirror state-action pairs together. The presence of a hashed cell indicates that it is not an exact homomorphism. The agent π solves the problem in this abstract domain.



2.42	**2.17**	■	4.16	4.64	5.16
2.70	3.01	3.35	3.74	■	4.64
2.42	2.70	■	3.35	3.74	4.16
■	2.42	2.17	■	3.35	■
1.94	2.17	2.42	2.70	3.01	**2.70**
1.74	1.94	■	2.42	2.70	■

Figure 3: The optimal values at each approachable cell. The bold-faced values are not exactly matched across the symmetry axis.

in the state s_{23}. In general, we get an *under-determined* set of equations for the action-value function at state s_{23} as

$$Q^*(s_{23},a)=\begin{cases}r_u+0.73\gamma p_u+3.01\gamma & \text{if } a=u\\ r_d-0.73\gamma p_d+3.74\gamma & \text{if } a=d.\end{cases} \tag{9}$$

In the original navigational grid-world problem $p_u = p_d = 1$, i.e. each action leads deterministically to the indented reachable cell, and $r_u = r_d = r_n$, where r_n is a fixed small negative reward. We can break the ≈MDP similarity by setting[7] $p_u = p_d := 0$, i.e. the actions behave in the opposite way in the lower half, $r_u := r_n + 0.73\gamma$ and $r_d := r_n - 0.73\gamma$, without changing the ≈Q-uniform similarity. In fact, we can have infinite combinations of rewards and transitions to get a set of modified domains since the set of equations (9) is under-determined.

This set of *modified* domains, by design, no longer allows the approximate MDP homomorphism of Figure 2. Every state is different in terms of reward and transition structure across the ≈MDP axis of Figure 1. Any one-step model similarity abstraction would be approximately of the same size as the original problem. However, if we consider Q-uniform homomorphisms, i.e. state-action pairs are merged if the action-values are close, then the set of modified domains has a same Q-uniform homomorphism.

In GRL, it is natural to assume that the (expected) rewards are function of realized history. The above modification argument is more likely to hold in a GRL setting: the reward and transition similarity might be hard to satisfy. Therefore, a GRL agent is better to consider such non-MDP homomorphisms to cover more domains with a single abstract model. Now we ask the main question, does such a non-MDP homomorphism, e.g. Q-uniform homomorphism, have a guaranteed solution for the original problem? In the next section, we answer this question in affirmative for Q-uniform homomorphisms (Theorem 8ii), but in negative for V-uniform homomorphisms with a weaker positive result (Theorem 10ii).

Key Elements to Go Beyond MDPs

This section introduces the key elements of the paper that enables us to prove performance bounds for non-MDP homomorphisms.

A Stochastic Inverse and Surrogate MDP

The key idea to get a near-optimal policy of the true environment P is to transform P_ψ into a surrogate MDP on the abstract state-action space. Afterwards, the optimal policy of this surrogate MDP is uplifted to P. This technique of casting a non-MDP process as an MDP has been used in ESA (Hutter 2016). To get this surrogate MDP, we define a stochastic inverse B of the homomorphism ψ as a probability measure over the history-action space given an abstract state-action pair, formally, $B : \mathcal{S} \times \mathcal{B} \to \Delta(\mathcal{H} \times \mathcal{A})$. Moreover, we require $B(ha|sb) := 0$ for any $\psi(h, a) \neq (s, b)$. The surrogate MDP is defined as

$$p_B(s'r'|sb) := \sum_{\tilde{h}\in\mathcal{H},\tilde{a}\in\mathcal{A}} P_\psi(s'r'|\tilde{h}\tilde{a})B(\tilde{h}\tilde{a}|sb). \tag{10}$$

It might seem like a paradoxical idea to solve a non-Markovian P_ψ using an MDP p_B, but the paradox is superficial. It is the stochastic inverse that complements the non-Markovianness of P_ψ. Finding such an inverse algorithmically, hence the surrogate MDP, is not a trivial task in general (Hutter 2016).

This action-dependent stochastic inverse separates our work from the action-independent weighting function considered by Abel, Hershkowitz, and Littman (2016). Although, learning of such weighting function is beyond the scope of this paper, an action-independent weighting function is not learnable. Because, when this weighting function is built from the true sampling distribution, it becomes action-dependent. Hutter (2016) constructs such a learnable action-dependent inverse for the state abstraction case. Fortunately, the choice of B becomes irrelevant in Q-uniform

[7] $p_u = 0$ implies that the action u now takes the agent to the down cell and vice versa for the action d.

homomorphisms (see Theorem 8), however this is not the case in V-uniform homomorphisms (see Theorem 10).

Similar to the original process, we also define the (action-)value functions for the surrogate MDP p_B on the abstract state-action space $\mathcal{S} \times \mathcal{B}$ with an abstract state-based policy π. The action-value function is given as

$$q^\pi(s,b) := \sum_{\tilde{s} \in \mathcal{S}, \tilde{r} \in \mathcal{R}} p_B(\tilde{s}\tilde{r}|sb)\left(\tilde{r} + \gamma v^\pi(\tilde{s})\right) \tag{11}$$

where the value function is

$$v^\pi(s) := \sum_{\tilde{b} \in \mathcal{B}} q^\pi(s,\tilde{b})\pi(\tilde{b}|s). \tag{12}$$

An abstract state-based *optimal* policy π^* is a value maximizer, i.e. $\pi^* \in \arg\max_{\tilde{\pi}} v^{\tilde{\pi}}$.

Representative Abstract Policy

As discussed earlier, we are primarily interested in the optimal policies of the surrogate MDP. However, it is also interesting to consider a general policy case (e.g. Theorems 4, 5, 7 and 9) akin to an on-policy result where we uplift a representative policy. We use any arbitrary member as a representative policy π_R on the abstract state s.

$$\pi_R(\cdot|s) := \Pi_\psi(\cdot|\tilde{h}), \quad \text{for some } \tilde{h} \in \psi^{-1}(s). \tag{13}$$

This arbitrary choice of representative introduces a policy representation error ε_Π for each abstract state s, expressed as

$$\varepsilon_\Pi(s) := \sup_{\tilde{h} \in \psi^{-1}(s)} \left\| \pi_R(\cdot|s) - \Pi_\psi(\cdot|\tilde{h}) \right\|_1. \tag{14}$$

This representation error is small/zero when the induced abstract policy Π_ψ is approximately/piecewise constant, i.e. $\Pi_\psi(\cdot|h) = \Pi_\psi(\cdot|h')$ for all $\psi(h) = \psi(h')$.

In the next section, we provide the main results of this work. We construct a near-optimal policy for the original process from the surrogate MDP even if the homomorphism is non-MDP.

Main Results

We analyze three types of homomorphisms in this work: MDP, Q-Uniform and V-Uniform homomorphisms[8]. Both Q and V-Uniform homomorphisms are non-Markovian by definition. In general, MDP and Q-Uniform homomorphisms admit a *deterministic* near-optimal policy of the original process, while V-Uniform homomorphisms do not.

Markov Decision Process Homomorphisms

A homomorphism is an MDP homomorphism if the induced abstract process P_ψ is an MDP. Then, there exists a process p_{MDP} such that for all $\psi(h,a) = (s,b)$ and for all $\tilde{s}$ and $\tilde{r}$, it holds:

$$P_\psi(\tilde{s}\tilde{r}|ha) = p_{\text{MDP}}(\tilde{s}\tilde{r}|sb). \tag{15}$$

Using the above condition in (10), renders $p_B = p_{\text{MDP}}$ and independent of B. The condition (15) is a *stronger* version of the bisimulation condition (Givan, Dean, and Greig 2003) that is generalized to joint history-action pairs. This condition is strong enough to preserve the optimal (action-)value functions of the original process (see Theorem 6). But, it is not strong enough to preserve arbitrary policy (action-)value functions (see Theorem 4). Unless we define a notion of action-value function representative and a corresponding representation error. For an abstract state-action pair, the representative action-value is defined as

$$Q^\Pi(\psi^{-1}(s,b)) := Q^\Pi(\tilde{h},\tilde{a}), \text{ for some } \psi(\tilde{h},\tilde{a}) = (s,b) \tag{16}$$

and the representation error of the action-value function is expressed as

$$\varepsilon_Q(s) := \sup_{\tilde{h},\tilde{a},\tilde{b}: \psi(\tilde{h},\tilde{a}) = (s,\tilde{b})} \left| Q^\Pi(\psi^{-1}(s,\tilde{b})) - Q^\Pi(\tilde{h},\tilde{a}) \right|. \tag{17}$$

Similar to ε_Π, this representation error is small/zero if the action-value function is approximately/piecewise constant. At this point, we have all the required components properly defined to state the first theorem of the paper.

Theorem 4. $(\psi_{\text{MDP}\Pi})$ *Let ψ be a homomorphism such that P_ψ is an MDP, then for any policy Π and all $\psi(h,a) = (s,b)$ it holds:*

$$\begin{aligned} \left|q^{\pi_R}(s,b) - Q^\Pi(h,a)\right| &\leq \frac{\gamma\varepsilon_{\max}}{1-\gamma} \quad \textit{and} \\ \left|v^{\pi_R}(s) - V^\Pi(h)\right| &\leq \frac{\varepsilon_{\max}}{1-\gamma} \end{aligned}$$

where $\varepsilon_{\max} := \max_{\tilde{s} \in \mathcal{S}} \left(\varepsilon_Q(\tilde{s}) + \frac{\varepsilon_\Pi(\tilde{s})}{1-\gamma}\right)$.

The above theorem shows that the surrogate MDP *approximately* preserves the (action-)value functions of the original process for any arbitrary policy. However, these (action-)value functions are preserved *exactly* if we further impose a policy uniformity condition in addition to an MDP assumption.

Theorem 5. $(\psi_{\text{MDP}\Pi=})$ *Let ψ be a homomorphism such that P_ψ is an MDP and $\Pi(\cdot|h) = \Pi(\cdot|h')$ (i.e. the policy similarity condition holds) for some policy Π and for all $\psi(h) = \psi(h')$. Then for all $\psi(h,a) = (s,b)$ it holds:*

$$q^{\pi_R}(s,b) = Q^\Pi(h,a) \quad \textit{and} \quad v^{\pi_R}(s) = V^\Pi(h).$$

Theorems 4 and 5 are important but not very useful results. As already discussed, we are interested in the (near-)optimal policies of the original process. And, we want to find the abstract policies that can be lifted with a performance guarantee from the abstract state-action space to the original history-action space.

Theorem 6. $(\psi_{\text{MDP}*})$ *Let ψ be a homomorphism such that P_ψ is an MDP, then for all $\psi(h,a) = (s,b)$ it holds:*

(i) $q^*(s,b) = Q^*(h,a)$ *and* $v^*(s) = V^*(h)$.

(ii) $V^*(h) = V^{\breve{\Pi}}(h)$
where $\breve{\Pi}(h) :\in \psi_s^{-1}\left(\pi^*(s)\right)$ *for any* $\psi(h) = s$.

[8] Due to the limited space, we omit the proofs. The proofs can be found in the extended version of this paper (Majeed and Hutter 2018).

For any MDP homomorphism, the performance guarantee is provided by Theorem 6ii. The abstract optimal policy π^* is also an optimal policy for the original process when lifted to the original history-action space.

Q-Uniform Homomorphisms

In this section, we relax the MDP condition (see Equation 15) on the abstract-process provided by the homomorphism. We show that there still exists an abstract policy that is near-optimal in the original process (see Theorem 8ii). We start with proving a value loss bound for an arbitrary policy when the action-value function of the original process is *approximately* ψ-uniform.

Theorem 7. (ψ_{Q^Π}) *Assume* $\left|Q^\Pi(h,a) - Q^\Pi(h',a')\right| \le \varepsilon$ *for some policy* Π *and for all* $\psi(h,a) = \psi(h',a')$*. Then for all* $\psi(h,a) = (s,b)$ *it holds:*

$$\begin{aligned} \left|Q^\Pi(h,a) - q^{\pi_R}(s,b)\right| &\le \varepsilon + \frac{\gamma\varepsilon(s)}{1-\gamma} \quad \textit{and} \\ \left|V^\Pi(h) - v^{\pi_R}(s)\right| &\le \frac{\varepsilon(s)}{1-\gamma} \end{aligned}$$

where $\varepsilon(s) := 2\varepsilon + \frac{\varepsilon_\Pi(s)}{1-\gamma}$*.*

The following theorem *improves* the optimal policy value loss bounds, *cf.* Theorem 7, and establishes the existence of a near-optimal policy of the original history-action space in the abstract state-action space.

Theorem 8. (ψ_{Q^*}) *Let* $|Q^*(h,a) - Q^*(h',a')| \le \varepsilon$ *for all* $\psi(h,a) = \psi(h',a')$*, then for all* $\psi(h,a) = (s,b)$ *it holds:*

(i) $|Q^*(h,a) - q^*(s,b)| \le \frac{2\varepsilon}{1-\gamma}$ *and*
$|V^*(h) - v^*(s)| \le \frac{2\varepsilon}{1-\gamma}$.

(ii) $0 \le V^*(h) - V^{\breve{\Pi}}(h) \le \frac{4\varepsilon}{(1-\gamma)^2}$
where $\breve{\Pi}(h) :\in \psi_s^{-1}\left(\pi^*(s)\right)$ *for any* $\psi(h) = s$*.*

It is important to note that Theorem 8 holds for any stochastic inverse B. Every choice of B gives a different surrogate MDP p_B, so the theorem provides a *near-optimal* performance guarantee for the *uplifted* abstract optimal policies of *any* possible surrogate MDP. Therefore, for any non-MDP Q-uniform homomorphism and a fixed B there exists an uplifted near-optimal policy ($\breve{\Pi}$ from Theorem 8ii).

V-Uniform Homomorphisms

All the previous results are valid for any choice of the stochastic inverse B. However, for V-uniform homomorphisms, the results are explicitly dependent on B (see Theorem 9 and 10). We need a couple of more entities to express the results of this section. We denote the B-average of the action-value function of the original process as

$$\langle Q^\Pi(\psi^{-1}(s,b))\rangle_B := \sum_{\tilde{h}\in\mathcal{H},\tilde{a}\in\mathcal{A}} Q^\Pi(\tilde{h},\tilde{a})B(\tilde{h}\tilde{a}|sb). \quad (18)$$

Furthermore, we can decompose B into two distinct parts: action dependent and independent. With an abuse of notation, assume an arbitrary joint distribution B over $\mathcal{H},\mathcal{A},\mathcal{S}$ and $\mathcal{B}$. By using the chain rule of probability distributions on B,

$$\begin{aligned} B(ha|sb) &= B(h|sb)B(a|bhs) \\ &= \frac{B(hs)B(b|hs)}{B(sb)}B(a|bhs) \\ &\overset{(a)}{=} \frac{B(hs)B(b|h)}{B(sb)}B(a|bh) \\ &= B(h|s)\frac{B(b|h)}{B(b|s)}B(a|bh) \\ &= \underbrace{B(h|s)}_{\text{action-independent}} \cdot \overbrace{\left(\frac{B(ab|h)}{B(b|s)}\right)}^{\text{action-dependent}} \quad (19) \end{aligned}$$

(a) follows from Assumption 3, the state is determined only by the history.

Using the action-dependent part from (19), we define a history and state based induced measure on the original action space for any B and an abstract state based policy π as

$$B^\pi(a|hs) := \sum_{\tilde{b}\in\mathcal{B}} \left(\frac{B(a\tilde{b}|h)}{B(\tilde{b}|s)}\right)\pi(\tilde{b}|s). \quad (20)$$

This seemingly complex and arbitrary relationship has a well-structured explanation. If *approximately*, the B distribution is linked to the actual dynamics of an agent π acting in the abstract state-action space, i.e. $B(b|s) \approx \pi(b|s)$, then $B^\pi(a|hs) \approx B(a|h)$, which is effectively a *shadow* agent induced by the agent π on the original history-action space.

To prove a result analogous to Theorem 7 for a V-uniform homomorphism, we need to impose an extra condition on B, *cf.* Theorem 7, which requires a structure on B and/or on the underlying original process. For general B, there exist some known counter examples (Hutter 2016).

Theorem 9. (ψ_{V^Π}) *Let* $\left|V^\Pi(h) - V^\Pi(h')\right| \le \varepsilon$ *for some policy* Π *and for all* $\psi(h) = \psi(h')$*, and* $\left|\sum_{\tilde{a}\in\mathcal{A}} Q^\Pi(h,\tilde{a})B^{\pi_R}(\tilde{a}|hs) - V^\Pi(h)\right| \le \varepsilon_B$ *for all* $s = \psi(h)$*, then it holds:*

$$\begin{aligned} \left|\langle Q^\Pi(\psi^{-1}(s,b))\rangle_B - q^{\pi_R}(s,b)\right| &\le \frac{\gamma(\varepsilon+\varepsilon_B)}{1-\gamma} \quad \textit{and} \\ \left|V^\Pi(h) - v^{\pi_R}(s)\right| &\le \frac{\varepsilon+\varepsilon_B}{1-\gamma}. \end{aligned}$$

In Theorem 7, we had an absolute loss-bound for action-value functions but in Theorem 9 we only have a B-average relationship. So far, we were able to get a near-optimal performance guarantee when the optimal policy of a surrogate MDP is uplifted to the original process (see Theorems 6ii and 8ii). However, there does not exist such a near-optimal performance guarantee for V-uniform homomorphisms. A *deterministic* abstract policy could be arbitrarily worse off when uplifted to the original process (Hutter 2016, Theorem 10) in V-uniform state-abstractions, which are a special case of V-uniform homomorphisms. However, a relatively weak result is still possible.

Theorem 10. (ψ_{V^*}) *Let* $|V^*(h) - V^*(h')| \leq \varepsilon$ *for all* $\psi(h) = \psi(h')$ *and* $|\sum_{\tilde{a} \in \mathcal{A}} Q^*(h, \tilde{a}) B^{\pi^*}(\tilde{a}|hs) - V^*(h)| \leq \varepsilon_B$ *for all* $s = \psi(h)$*, then for all* $\psi(h, a) = (s, b)$ *it holds:*

(i) $|\langle Q^*(\psi^{-1}(s, b))\rangle_B - q^*(s, b)| \leq \frac{3\gamma(\varepsilon+\varepsilon_B)}{(1-\gamma)^2}$ *and* $|V^*(h) - v^*(s)| \leq \frac{3(\varepsilon+\varepsilon_B)}{(1-\gamma)^2}$.

(ii) If $\varepsilon + \varepsilon_B = 0$ *then* $\psi(h, \Pi^*(h)) = (s, \pi^*(s))$ *for all* $\psi(h) = s$.

In the *approximate* case, i.e. $\varepsilon + \varepsilon_B > 0$, Theorem 10 is not as useful as Theorem 8 because of the missing performance guarantee, *cf.* Theorem 8ii. However, it is still an important theorem for the *exact* V-uniform homomorphisms, i.e. $\varepsilon + \varepsilon_B = 0$. In that case, it is guaranteed that the optimal actions of *all* member histories are mapped to the *same* abstract optimal action (see Theorem 10ii).

Discussion, Outlook and Conclusion

In this paper we analyzed *approximate* homomorphisms of a general history-based environment. The main idea was to find a *deterministic* policy in the abstract state-action space such that, when uplifted, it is a near-optimal policy in the original problem. Using the surrogate MDP technique, we proved near-optimal performance bounds for both MDP (Theorem 6ii) and Q-uniform homomorphisms (Theorem 8ii). In general, there does not exist a near-optimal *deterministic* uplifted policy for V-uniform homomorphisms. However, we proved a weaker result (Theorem 10ii) for the *exact* V-uniform homomorphisms: the optimal actions of the member histories are mapped to the same abstract optimal action at the corresponding state of the surrogate MDP.

Versus ESA. We borrow some notation and techniques from Hutter (2016). But this work is crucially different from ESA. Apart from the obvious difference of being a generalization to homomorphisms, there are also some other key differences. In ESA, the policy Π is required to be state uniform for various of the main results (Hutter 2016, Theorems 1,5,6 and 9), whereas we do not make any such assumption. The extra conditions on Theorems 9 and 10 are *weaker* than the policy-uniformity condition, *cf.* (Hutter 2016, Theorems 6 and 9), and do not have direct counterparts in ESA.

Versus Options. As briefly addressed in the introduction section, the options framework does not have any provable performance guarantees, yet. Whereas our restriction of uplifting a state-based policy and using a deceptively "spatial-looking" abstraction of actions have such guarantees. Since we allow the action mapping part of ψ to be a function of history, which is arguably a function of time, our framework also admits temporal dependencies. It enables ψ to model much more than mere renaming of the original action space distributions. A thorough comparison between these two approaches is left for future work.

Outlook

The results in this work are quite general but there are various open questions left for future research.

Reinforcement Learning (Learning Problem). For a given homomorphism ψ, the most obvious question we left open is the choice of B. We call this the *learning problem.* Two of the three main results in this work (Theorems 6ii and 8ii) are valid for any choice of B, so any fixed B would suffice. But the third main result (Theorem 10ii) is very much involved with the choice of B. However, it is not a strong result in itself. Nevertheless, in a state-abstraction context, B facilitates learning of the surrogate MDP from the induced abstract process (Hutter 2016). Therefore, it is an intriguing direction to explore for homomorphisms.

Feature Reinforcement Learning (Discovery Problem). The focus of this paper is to provide performance guarantees for a given homomorphism. While in practice, the agent has to learn such a reduction/model from experience. It is known as the *discovery problem* (Li, Walsh, and Littman 2006) in RL and *Feature Reinforcement Learning* (FRL) (Hutter 2009) in the GRL context. It is non-trivial to solve this problem even in a state-abstraction framework (Hutter 2016). Our result can help to build such an FRL algorithm for homomorphisms, e.g. during the model learning/building, the algorithm may use the bounds from this work to select/discard a candidate model.

Special Environment Classes. In general, we do not use/exploit structure of the underlying original process. However, effects of a specific model class can be expressed in terms of the (action-)value functions. For example, if the original process is a finite state POMDP then our results provide the performance-loss guarantee by representing a belief-state based value function of the POMDP by a state-based value function. A similar argument can be rendered for various other types of model classes. Since the results in this work are general, they are not expected to gracefully scale down to some class specific tight performance bounds. Nevertheless, it is an important agenda to get the scaled-down variants of these results for some specific model classes.

Continuous state-action space. The results in this paper easily extend to the continuous state-action space homomorphisms for *measurable* maps. The summations change to integrals and the measurability constraint make sure that these integrals are well-defined. In this case, a homomorphism map has a natural interpretation of being a discretization of the underlying space. However, it is sometimes desirable to use a restricted continuity condition, e.g. Lipschitz or Holder continuity, rather than the weak measurability constraint. A continuous state-action homomorphism under some restricted continuity constraints would be a nice generalization of our results.

Fully Generalized Homomorphism. In a sense our results are not *fully* general since we assumed a structure on the homomorphism. A fully generalized homomorphism formulation with no $\psi(h, a) = (f(h), b)$ assumption would be an interesting extension of this work. However, lifting this condition may lead to some bizarre non-causal effects, e.g. the *current* abstract state would be decided by the *next* action!

Conclusion

In conclusion, our results show that in GRL the near-optimal performance guarantee is not limited only to MDP homomorphisms. It is sometimes possible to have non-MDP models, i.e. Q-uniform homomorphisms, with bounded performance loss guarantees.

Acknowledgements. We thank *Elliot Catt* and *Amy Zhang* for proofreading earlier drafts and the anonymous reviewers for their valuable feedbacks. This work has in parts been supported by Australian Research Council grant DP150104590.

References

Abel, D.; Hershkowitz, D. E.; and Littman, M. L. 2016. Near Optimal Behavior via Approximate State Abstraction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2915–2923.

Bellman, R. 1957. A Markovian Decision Process. *Indiana University Mathematics Journal* 6(4):679–684.

Givan, R.; Dean, T.; and Greig, M. 2003. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence* 147(1-2):163–223.

Hutter, M. 2009. Feature Reinforcement Learning: Part I. Unstructured MDPs. *Journal of Artificial General Intelligence* 1(1):3–24.

Hutter, M. 2016. Extreme State Aggregation Beyond Markov Decision Processes. *Theoretical Computer Science* 650:73–91.

Lattimore, T., and Hutter, M. 2014. Near-optimal PAC bounds for discounted MDPs. *Theoretical Computer Science* 558(C):125–143.

Leike, J. 2016. *Nonparametric General Reinforcement Learning*. Ph.D. Dissertation, Australian National University.

Li, L.; Walsh, T. J.; and Littman, M. L. 2006. Towards a Unified Theory of State Abstraction for MDPs. *In Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics* 531–539.

Littman, M. L.; Dean, T. L.; and Kaelbling, L. P. 1995. On the Complexity of Solving Markov Decision Problems. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 394–402.

Majeed, S. J., and Hutter, M. 2018. Performance Guarantees for Homomorphisms Beyond Markov Decision Processes.

McGovern, A.; Sutton, R. S.; and Fagg, A. H. 1997. Roles of Macro-Actions in Accelerating Reinforcement Learning. In *incompleteideas.net*, 13–18.

Puterman, M. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Ravindran, B., and Barto, a. G. 2003. Relativized Options: Choosing the Right Transformation. *Proceedings of the Twentieth International Conference on Machine Learning* 608–615.

Ravindran, B., and Barto, A. G. 2004. Approximate Homomorphisms: A Framework for Non-exact Minimization in Markov Decision Processes. *Proceedings of the Fifth International Conference on Knowledge Based Computer Systems (KBCS 04)* 19–22.

Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement Learning in Finite MDPs : PAC Analysis. *Journal of Machine Learning Research* 10:2413–2444.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement Learning: An Introduction Second*. MIT press Cambridge, 2nd edition.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1):181–211.

Taylor, J.; Precup, D.; and Panangaden, P. 2008. Bounding Performance Loss in Approximate MDP Homomorphisms. *Advances in Neural Information Processing Systems (NIPS) 21* 1649–1656.

Whitt, W. 1978. Approximations of Dynamic Programs, I. *Mathematics of Operations Research* 3(3):231–243.

Wolfe, A. 2010. Paying Attention To What Matters : Observation Abstraction In Partially Observable Environments. *Open Access Dissertations* (February 2010):188.