We here include proofs omitted from the paper.

**Theorem 3.1** (Efficient Abstractions). Consider any transitive predicate on state pairs, \( p \), that takes computational complexity \( c_p \) to evaluate for a given state pair. The state abstraction type \( \phi_p \) that induces the smallest abstract state space can be computed\(^\dag\) in \( O(|S|^2 \cdot c_p) \).

**Proof.** Let \( c_p \) denote the computational complexity associated with computing the predicate \( p \) for a given state pair. Consider the algorithm consisting of the following four rules for constructing abstract clusters (which define the abstract states) using queries to each of the \(|S|^2\) state pairs. Let \((s_i, s_j)\) denote the current state pair:

1. If \( p(s_i, s_j) \) is true, and neither state is in an abstract cluster yet, make a new cluster consisting of these two states.
2. If \( p(s_i, s_j) \) is true and only one of the states is already in a cluster, add the other state to the existing cluster.
3. If \( p(s_i, s_j) \) is true and both \( s_i \) and \( s_j \) are in different clusters, merge the clusters.
4. If \( p(s_i, s_j) \) is false, add each state not yet in a cluster to its own cluster.

Running this algorithm makes one query per state pair, of which there are \(|S|^2\). Thus, the complexity is \( O(|S|^2 \cdot c_p) \).

From steps 1-3, after iterating through the possible state pairs, there cannot exist a state pair \((s_x, s_y)\) such that \( p(s_x, s_y) \) is true but \( s_x \) and \( s_y \) are in different clusters. Further, by transitivity, when we apply the cluster merge in step 3, we are guaranteed that every state pair in the resulting cluster necessarily satisfies the predicate. Thus, we compute the smallest clustering definable by \( p \). \(\square\)

**Theorem 3.2**. The \( \phi_{Q^*_d} \) abstraction type is a subclass of \( \phi_{Q^*_\varepsilon} \), studied by Abel et al. (2016) and Hutter (2016), with \( d = \varepsilon \), and therefore, for a single MDP:

\[
V^*(s_0) - V^{\pi_{Q^*_d}}(s_0) \leq \frac{2d\text{RMAX}}{(1 - \gamma)^2}.
\]  

**Proof.** For any two state-action pairs that satisfy the predicate \( \phi_{Q^*_d} \), we know by definition of the predicate that for each action \( a \), there exists a \( Q_{\text{lower}} \) such that:

\[
Q_{\text{lower}} \leq Q(s_1, a) \leq Q_{\text{lower}} + d,
\]

\[
Q_{\text{lower}} \leq Q(s_2, a) \leq Q_{\text{lower}} + d.
\]

Therefore, for each action \( a \):

\[
|Q(s_1, a) - Q(s_2, a)| \leq d.
\]  

Therefore, \( \phi_{Q^*_d} \) is a subclass of \( \phi_{Q^*_\varepsilon} \). \(\square\)

**Theorem 3.2** (Abstract State Space Size). For a given \( d \), the function belonging to the transitive abstraction type \( \phi_{Q^*_d} \) that induces the smallest possible abstract state space size is at most \( 2^{|A|} \) times larger than that of the maximally compressing instance of type \( \phi_{Q^*_\varepsilon} \), for \( d = \varepsilon \). Thus, letting \( S_d \) denote the abstract state space associated with the maximally compressing \( \phi_{Q^*_\varepsilon} \), and letting \( S_\varepsilon \) denote the abstract state space associated with the maximally compressing \( \phi_{Q^*_\varepsilon} \):

\[
|S_\varepsilon| \cdot 2^{|A|} \geq |S_d|.
\]

**Proof.** Let \( M \) be an arbitrary MDP. Consider a set of states \( \tilde{S} \subseteq S \) clustered together under \( \phi_{Q^*_d} \), and, in particular, consider the \( Q \)-values of all states in \( \tilde{S} \) for a particular action, \( a \in A \). Note that, by construction of \( \phi_{Q^*_d} \), for any \( s, s' \in \tilde{S} \):

\[
\forall s, s' \in \tilde{S} : |Q(s, a) - Q(s', a)| \leq \varepsilon,
\]

Recall that, intuitively, \( \phi_{Q^*_d} \) is a discretization of the interval \([0, \text{VMAX}]\) where \( d \) controls the placement of boundaries, forming buckets of \( Q \)-values. The \( Q \)-values for all states in \( \tilde{S} \) and for action \( a \) reside in a single sub-interval of length \( \varepsilon \).

Letting \( d = \varepsilon \), the placement of boundaries that form \( \phi_{Q^*_d} \) could break the \( \varepsilon \)-interval of \( Q \)-values for the non-transitive
cluster $\hat{S}$ no more than once, resulting in the creation of at most two new state clusters in $\hat{\phi}_{Q_2}$. Repeating the process $\forall a \in A$, these separations within the original cluster compound, resulting in at most $2^{|A|}$ such subdivisions and, accordingly, $2^{|A|}$ clusters in $\hat{\phi}_{Q_2}$ for each cluster in $\hat{\phi}_{Q_2}$.

**Corollary 3.3.1 (PAC Value Loss).** Consider any state-abstraction type $\phi_p$ with value loss $\tau_p$, that is, in the traditional single task setting:

$$\forall_{s \in S} : V^*_M(s) - V^p_M(s) \leq \tau_p.$$  \hfill (4)

Then, the PAC abstraction $\hat{\phi}_p^\delta$, in the lifelong setting, has value loss:

$$\mathbb{E}_{M \sim D} \left[ V^*_M(s) - \forall_{s \in S} : V^p_M(s) \right] \leq \varepsilon (1 - 3\delta) \tau_p + 3\delta \text{VMAX}. \hfill (5)$$

**Proof.** By definition of PAC abstractions, with probability $1 - \delta$, the abstraction function $\phi_p^\delta \in \phi_p$ aggregates iff $\rho_{\delta+\varepsilon}$ for some small $\varepsilon \in (-\delta, \delta)$.

Then, with probability $1 - \delta$, there is at least a $1 - \delta - \varepsilon$ chance that the predicate holds for a particular state, by definition of $\rho_p^\delta$. Thus, by definition of $\rho_p^\delta$, with probability $(1 - \delta)(1 - \delta - \varepsilon)$, the state abstraction correctly aggregates, and consequently the inherited value loss $\tau_p$ bound holds. If the abstraction incorrectly aggregates, the value loss can be up to VMAX.

Letting $\varepsilon = \delta$, we see that the PAC loss is at worst upper bounded by a convex mixture of $\tau_p$ with probability $(1 - 3\delta)$, and with probability $3\delta$, is VMAX. Thus, the value loss of $\phi_p^\delta$ is:

$$\forall_{s \in S} : \mathbb{E}_{M \sim D} \left[ V^*_M(s) - V^p_M(s) \right] \leq \varepsilon (1 - 3\delta) \tau_p + 3\delta \text{VMAX}. \hfill (6)$$

**Theorem 3.4 (PAC Abstraction Sample Bound).** Let $\mathcal{A}_p$ be an algorithm that given an MDP $M = (S, A, R, T, \gamma)$ as input can determine if $p(s_1, s_2)$ is true for any pair of states, for any state abstraction type.

Then, for a given $\delta \in (0, 1]$ and $\varepsilon \in (-\delta, \delta)$, we can compute $\phi_p^{\delta+\varepsilon}$ after $m \geq \frac{\ln \frac{2}{\delta}}{\varepsilon^2}$ sampled MDPs from $D$.

**Proof.** We are given as input a $\delta \in (0, 1]$, a distribution over MDPs $D$, and the algorithm $\mathcal{A}_p$, which, given an MDP $M$ and a state pair outputs $p_M(s, s')$.

Consider an arbitrary pair of states $s$ and $s'$. For $m$ sampled MDPs, the algorithm $\mathcal{A}_p$ can produce a sequence of $m$ predicate evaluations:

$$p_1(s, s'), \ldots, p_m(s, s'). \hfill (7)$$

Let $\bar{p}$ be the empirical mean over the predicate sequence:

$$\bar{p} = \frac{1}{m} \sum_{i=1}^{m} p_i(s, s'). \hfill (8)$$

The clustering algorithm is quite simple: for our input $\delta \in (0, 1]$, cluster all state pairs $(s, s')$ such that $\bar{p}(s, s') \geq 1 - \delta$ after $m$ samples.

We now prove that, for a particular setting of $m$, the resulting cluster assignments constitute a state abstraction that clusters a pair of states only if the predicate is true with high probability.

First, let $\bar{p}$ denote the probability that $p$ is true over the distribution:

$$\bar{p}(s, s') = \Pr_{M \sim D} \{ p(s, s') = 1 \}. \hfill (9)$$

Using Hoeffding's bound, we upper bound the probability that $\bar{p}$ deviates from $\bar{p}$ by more than some small $\varepsilon \in (0, \delta)$:

$$\Pr \{ |\bar{p}(s, s') - \bar{p}(s, s')| \geq \varepsilon \} \leq 2e^{-2m\varepsilon^2}. \hfill (10)$$

Thus, for $\delta = 2e^{-2m\varepsilon^2}$:

$$\Pr \{ |\bar{p}(s, s') - \bar{p}(s, s')| < \varepsilon \} > 1 - \delta. \hfill (11)$$

Rewriting:

$$\Pr \{ |\bar{p}(s, s') - \bar{p}(s, s')| < \varepsilon \} > 1 - \delta \hfill (12)$$

By algebra, note that, when $m \geq \frac{\ln \frac{2}{\delta}}{\varepsilon^2}$, the condition of Equation 12 holds.

Let $\rho_p^\delta$ denote the predicate that is true if and only if $p$ is true over the distribution with high probability for a given $\delta$:

$$\rho_p^\delta(s_1, s_2) = \begin{cases} 
1 & \bar{p} \geq 1 - \delta \\
0 & \text{otherwise}.
\end{cases} \hfill (13)$$

Now, we form our state abstraction under the following rule:

$$\hat{\phi}_p^\delta(s_1) = \hat{\phi}_p^\delta(s_2) \equiv \bar{p}(s, s') > 1 - \delta. \hfill (14)$$

If, after $m$ samples, $\bar{p}$ were identical to $\bar{p}$, then we would have:

$$\forall_{s, s'} : \Pr_{M \sim D} \{ \rho_p^\delta(s, s') \leq \hat{\phi}_p^\delta(s_1) = \hat{\phi}_p^\delta(s_2) \} \geq 1 - \delta. \hfill (15)$$
Hence, $\rho$ deviates from $\overline{p}$ by at most $\varepsilon$ with probability $1 - \delta$. Thus, for some $\varepsilon \in (-\delta, \delta)$, $\rho + \varepsilon = \overline{p}$. Therefore, the clustering rule defined by Equation 16 ensures there exists an $\varepsilon$ such that, with high probability, we cluster according to:

$$\forall s_1, s_2 : p^\delta_{s_1 + \varepsilon}(s_1, s_2) \Leftrightarrow \phi^\delta_p(s_1) = \phi^\delta_p(s_2).$$

(18)

We conclude that, for $m \geq \frac{\ln \frac{2}{\varepsilon}}{\varepsilon^2}$ sampled and solved MDPs, we compute a lifelong PAC state abstraction $\phi^\delta_p$. \hfill $\square$

**Theorem 3.5.** Consider an MDP $M$ and an instance of R-Max (Brafman & Tennenholtz, 2002) that breaks ties using round-robin selection over actions. This algorithm is PAC-MDP in the raw state space. Next, pair a domain with any state-abstraction function $\phi$. If R-Max interacts with $M$ by projecting any received state $s$ through $\phi$, then R-Max is no longer guaranteed to be PAC-MDP in $M$. In fact, the number of mistakes made by R-Max can be arbitrarily large.

**Proof.** Consider the simple three state chain:

![Diagram of three state chain](image)

The agent has three actions, left, right, and loop, associated with their natural effects (left in $s_0$ is a self loop with reward 0, while right moves the agent to $s_1$, and so on).

In states $s_0$ and $s_1$, let the reward for loop be some small constant $\kappa$, and let the loop action in $s_3$ yield RMAX reward.

Let $\varepsilon = 0.1$, $\gamma = 0.95$, $s_0$ define the initial state, and $\kappa = 0.001$. Then

$$\forall \varepsilon \in \{s_0, s_1, s_2\} : \max_{a_1, a_2} Q^*_{\pi}(s, a_1) - Q^*_{\pi}(s, a_2) \leq \varepsilon.$$

Therefore, for $\varepsilon = 0.1$, a valid clustering assigns $\phi(s_0) = \phi(s_1)$. The R-Max knownness parameter for a state-action pair is given as $m$.

To break ties, we suppose R-Max chooses actions according to a round-robin policy, starting with action left. Thus, in the abstract, R-Max first chooses left, then right, then self loop, then left, right, self loop, and so on, until each state-action pair is known.

In the above problem, this sequence of actions will never lead the agent out of state $s_0$ or $s_1$. Therefore, after $m$ executions of these three actions across states $s_0$ and $s_1$, R-Max with $\phi$ will compute a transition model that never includes the ability to transition to $g$. Further, the action loop will have the largest reward associated with it—$\kappa$, a reward chosen to be arbitrarily small—which is thus arbitrarily worse than the goal reward. So, R-Max will make an unbounded number of mistakes. \hfill $\square$

**Corollary 3.5.1.** For any RL algorithm $\mathcal{A}$ whose policy updates during learning and an arbitrary state abstraction $\phi$.

Let $\mathcal{A}_\phi$ denote the algorithm yielded by projecting all incoming states to $\phi(s)$ before presenting them to $\mathcal{A}$, and let $M_\phi = \langle S_\phi, A, T_\phi, R_\phi, \gamma \rangle$, denote the abstract MDP induced by $\phi$ on $M$, where:

- $S_{\phi} = \{ \phi(s) : \forall s \in S \}$,
- $R_{\phi}(s, a) = \sum_{g \in \phi^{-1}(\phi(s))} w(g) R(g, a)$,
- $T_{\phi}(s, a, s') = \sum_{g \in G(s)} \sum_{g' \in G(s')} T(g, a, g') w(g)$,

with $w(s)$ is a fixed weighting function and $G(s) = \phi^{-1}(\phi(s))$. That is, $G(s)$ gets all of the true environmental states in the same cluster as $s$.

The process yielded by $\mathcal{A}_\phi$ interacting with $M$ is not identical to $\mathcal{A}$ interacting with $M_\phi$. That is, the expected trajectory taken by the agent is not the same in the two situations. Formally:

$$E_{\mathcal{A}}[s_t | s_0, \pi] \neq E_{\mathcal{A}_\phi}[s_t | s_0, \pi],$$

(19)

where $s_t$ is the state the agent arrives in after $t$ time steps.

**Proof.** Note that when $M_\phi$ is computed directly, the functions $R_\phi$ and $T_\phi$ assume a fixed weighting function $w(s)$.

Again consider the three state chain from the previous proof.

During typical interaction between $M$ and $\mathcal{A}_\phi$, however, no such fixed weighting function exists for any algorithm $\mathcal{A}$ that updates its policy. That is, the distribution of states the agent finds itself in will change as its policy changes, and therefore, $w(s)$ must change, too.

Thus, the process of $\mathcal{A}_\phi$ interacting with $M$ induces a sequence of interactions with abstract MDPs whose transition and rewards change along with the policy the agent follows. Thus, for any non-identity $\phi$, for any algorithm $\mathcal{A}$ whose policy changes over time, the resulting interaction is non-identical. \hfill $\square$