Notes on State Abstractions

Nan Jiang

September 28, 2018

A common aspect of methods surveyed in the previous lectures is that the size of dataset (or *sample size*) required to yield learning guarantees is polynomial in the size of the state space. When the size of the state space is very large, as will be the case in many challenging problems, the agent needs to generalize what is learned about one state to other states using prior knowledge.

One of the easiest-to-deploy generalization schemes is state abstraction (sometimes also called state aggregation/compression). A state abstraction is a mapping ϕ that maps the original (or primitive/raw) state space $\mathcal{S}$ to some finite *abstract* state space; for brevity we use $\phi(\mathcal{S})$ to denote the codomain of the mapping. Intuitively, if $s^{(1)}$ and $s^{(2)}$ are mapped to the same element, that is $\phi(s^{(1)}) = \phi(s^{(2)})$, they are treated as the same state.

Given a problem with state space $\mathcal{S}$ and an abstraction ϕ, a typical usage of ϕ is to convert every state s in the dataset D into $\phi(s)$, and run any tabular algorithm over D with the understanding that the state space is $\phi(\mathcal{S})$. For example, if we collect a dataset that consists of tuples (s, a, r, s'), we can view each tuple now as $(\phi(s), a, r, \phi(s'))$, and build a certainty-equivalence model over state space $\phi(\mathcal{S})$. This is always doable, although there might not be a well-defined MDP with state space $\phi(\mathcal{S})$ that is the groundtruth process for the converted dataset.

An obvious benefit of using state abstraction is the increase of effective sample size. Suppose we collected a dataset with n samples per (s, a) pair, and an abstraction ϕ maps $s^{(1)}$ and $s^{(2)}$ to the same abstract state x. Then, after applying the abstraction, we get $2n$ samples for the state-action pairs (x, a). In certainty-equivalence, we essentially double the sample size for estimating the transition and reward functions for a state-action pair and can enjoy lower *estimation errors*.

This advantage, of course, comes with a caveat, otherwise we could simply map every $s \in \mathcal{S}$ to the same abstract state and maximize the number of samples per state. The caveat is that if we aggregate states that are very different from each other, then we do not converge to the optimal policy even in the limit of infinite data; in other words, we may incur high *approximation errors*. The trade-off between approximation error and estimation error is a constant theme of statistical machine learning [1].

Intuitively, the approximation error is high when we aggregate states that are very different from each other. The question is, how should we define an (approximate) equivalence notion among states? Whether they share the same optimal action? Whether they share the same $Q^\star$ values? Whether they yield the same rewards and next-state distributions? It turns out that, these criteria define a hierarchy of different state abstractions, from lenient to strict. Lenient notions of abstractions yield more generalization benefits, but may not work well with certain algorithms; strict notions of abstractions preserve many properties of the original MDP and hence work with a wider range of algorithms, but the generalization benefits are relatively limited.

1 Exact abstractions

Below we define the hierarchy of abstractions. We start by considering *exact* abstractions that only aggregate states that are strictly equivalent. In the next section we will relax the condition and allow approximate equivalence.

Definition 1 (Abstraction hierarchy [2])**.** Given MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and state abstraction ϕ that operates on $\mathcal{S}$, define the following types of abstractions:

1. ϕ is $\pi^\star$-irrelevant if there exists an optimal policy $\pi^\star$, such that $\forall s^{(1)}, s^{(2)} \in \mathcal{S}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\pi^\star_M(s^{(1)}) = \pi^\star_M(s^{(2)})$.
2. ϕ is $Q^\star$-irrelevant if $\forall s^{(1)}, s^{(2)}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\forall a \in \mathcal{A}$, $Q^\star_M(s^{(1)}, a) = Q^\star_M(s^{(2)}, a)$.
3. ϕ is model-irrelevant if $\forall s^{(1)}, s^{(2)}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\forall a \in \mathcal{A}, x' \in \phi(\mathcal{S})$,

$$R(s^{(1)}, a) = R(s^{(2)}, a), \quad \sum_{s' \in \phi^{-1}(x')} P(s'|s^{(1)}, a) = \sum_{s' \in \phi^{-1}(x')} P(s'|s^{(2)}, a). \tag{1}$$

   Note that the condition on transition dynamics is essentially $P(x'|s^{(1)}, a) = P(x'|s^{(2)}, a)$. It will also be convenient to define a $|\phi(\mathcal{S})| \times |\mathcal{S}|$ matrix Φ, where

$$\Phi(x, s) = \mathbb{I}[\phi(s) = x].$$

   So $\Phi P(s, a)$ collapses the transition distribution over $\mathcal{S}$ to a distribution over $\phi(\mathcal{S})$, and the condition on transition dynamics can be rewritten as: $\Phi P(s^{(1)}, a) = \Phi P(s^{(2)}, a)$.

The following property of the hierarchy shows that $\pi^\star$-irrelevance is the most lenient and model-irrelevance is the most strict.

Theorem 1 (Theorem 2 of [2][1])**.** *Model-irrelevance implies $Q^\star$-irrelevance, which further implies $\pi^\star$-irrelevance.*

The proof is deferred to Section 2.

1.1 Model-irrelevance

Among the 3 notions of abstractions, $Q^\star$-irrelevance and $\pi^\star$-irrelevance are relatively straightforward. The definition of model-irrelevance, however, can be a litte counter-intuitive. So in this subsection we will try to develop this notion from scratch and see why the existing definition makes sense.

A naive condition A criterion that we may naturally come up with is the following: $\forall s^{(1)}, s^{(2)}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\forall a \in \mathcal{A}, s' \in \mathcal{S}$,

$$R(s^{(1)}, a) = R(s^{(2)}, a), \quad P(s'|s^{(1)}, a) = P(s'|s^{(2)}, a). \tag{2}$$

So essentially we ask that $s^{(1)}$ and $s^{(2)}$ to be aggregated only when they have the same rewards and the same transition distributions *over raw states* for all actions. This makes intuitive sense and in fact it is a sufficient condition for Eq.(1). So why don't we use this criterion instead?

[1] Li et al. [2] also included two additional types of abstractions as well as the raw state representation in the hierarchy theorem, which are omitted here.

The problem is that the criterion is too strict. Here is an example: Let M be an MDP with $x \in \mathcal{X}$ being its state, and C be a Markov chain with $z \in \mathcal{Z}$ being its state. Now we augment M into a new MDP M' with state space $\mathcal{X} \times \mathcal{Z}$, and we use (x, z) to denote its state. The reward function of M' is the same as M, in the sense that the z component of the augmented state is ignored. For transition, the x component evolves according to the transition dynamics of M (which depends on actions), and the z component evolves according to that of C.

It is obvious from the construction that M' is essentially equivalent to M and the z component is useless and should be ignored. In other words, a state abstraction $\phi : (x, z) \mapsto x$ is a perfect choice for M' and we want our model-irrelevance condition to accommodate ϕ. Unfortunately, upon examining Eq.(2) for ϕ, we find that as long as C has state-dependent transitions (i.e., C is a not an i.i.d. process), ϕ will not satisfy Eq.(2) in general. In contrast, such an example is perfectly accomodated by Eq.(1), since it compares the distributions from two state action pairs *after* integrating out the factor that is believed to be irrelevant.

Bisimulation can handle irrelevant factors in more general situations than the above example: Even if z' depends on z, x, a, and even x', as long as x's evolution does not depend on z and reward only depends on x and a, ignoring z yields bisimulation. (In fact this factorization view is both a sufficient and necessary condition, therefore an alternative definition, of bisimulation: I claim that we can always induce such a factorization of raw state representation from a bisimulation ϕ, such that ϕ simply drops the irrelvant part. Why? Hint: $\phi(s)$ is the relevant factor, and the remaining information in s is the irrelevant factor; you can verify from Eq.(1).)

Bibliographical Remarks and Discussions In RL literature, the notion of model-irrelevance was originally introduced as *bisimulations* [3], and was later generalized to *MDP homomorphisms* to handle action aggregation and permutation [4].

While model-irrelevance is the most strict notion in the abstraction hierarchy, it also secures the success of almost any tabular RL algorithm: given model-irrelevant ϕ, it is fundamentally impossible to distinguish between two datasets, one drawn from an abstract MDP M_ϕ that is a perfect compression of the original MDP M (the spec of this MDP is implicitly given in Equation 1), and the other drawn from the original MDP and converted using ϕ ($(s, a, r, s') \to (\phi(s), a, r, \phi(s'))$); for the purpose of analysis we can simply treat the algorithm as if it were run in M_ϕ, and any guarantee for the algorithm automatically extends.

On the other hand, if ϕ is $Q^\star$-irrelevant, the compression does not preserve rewards or dynamics in general. While some tabular algorithms can still be applied and their guarantees extend (e.g., Q-learning), these extensions are not automatic and need new analyses (see e.g., Section 8.2.3 in [5]). When it comes to $\pi^\star$-irrelevance, it is known that value-based and model-based algorithms may break down [6]; only policy search methods which directly optimize the return over a policy class can retain guarantees due to their robustness to agnosticity [7].

A couple of final remarks:

- Given the MDP specification, there is a bisimulation that is a common refinement of all other bisimulations (i.e., a minimal bisimulation uniquely exists; see proof at the end of this section), but finding it is an NP-hard computation problem [3] (and so is finding the minimal ϵ-approximate bisimulation, which will be introduced in the next section).

- Abstractions are also useful in planning since it reduces the effective size of the state space and hence computational complexity. In this case, model-irrelevance has an advantage over the other two types of abstractions, since it can be checked *without* performing planning. See [8] for an analogy of this situation in the learning setting.
- There is an RKHS view of Eq.(1): An abstraction ϕ induces a space of piece-wise constant functions over $\mathcal{S}$, denoted as $\mathcal{F}^\phi$. Two states are equivalent if their transition distributions (for each action) have the same *RKHS embeddings* in $\mathcal{F}^\phi$; in other words, a kernel two sample test using $\mathcal{F}^\phi$ will not be able to detect the difference between the two distributions [9].
- Our next topic will be fitted value iteration with generic function approximator, and both tabular methods and abstractions are its special cases. There we will see that $Q^\star$-irrelevance corresponds to a standard, supervised learning-type realizability assumption, and model-irrelevance implies *closedness* of the function class under Bellman operator.

Uniqueness of coarsest bisimulation

Proof. We prove by showing that for any bisimulations ϕ_1 and ϕ_2 of M, their *common coarsening* is also a bisimulation, denoted as ϕ_{12}. We define ϕ_{12} by giving its equivalence criterion: for any $s^{(1)}$ and $s^{(2)}$, $\phi_{12}(s^{(1)}) = \phi_{12}(s^{(1)})$ if and only if the two states are equivalent under either ϕ_1 or ϕ_2. Now we verify that ϕ_{12} is bisimulation. The reward condition is obviously satisfied, so it remains to check the transition condition.

Due to symmetry we consider any two states such that $\phi_1(s^{(1)}) = \phi_1(s^{(2)})$. For any $y' \in \phi_{12}(\mathcal{S})$,

$$
\begin{aligned}
P(y'|s^{(1)},a) &= \sum_{s'\in\phi_{12}^{-1}(y')} P(s'|s^{(1)},a) \\
&= \sum_{x'\in\phi_1(\phi_{12}^{-1}(y'))} \sum_{s'\in\phi_1^{-1}(x')\bigcap\phi_{12}^{-1}(y')} P(s'|s^{(1)},a) \\
&= \sum_{x'\in\phi_1(\phi_{12}^{-1}(y'))} \sum_{s'\in\phi_1^{-1}(x')} P(s'|s^{(1)},a) \qquad (\phi_1^{-1}(x') \text{ is always entirely inside } \phi_{12}^{-1}(y')) \\
&= \sum_{x'\in\phi_1(\phi_{12}^{-1}(y'))} \sum_{s'\in\phi_1^{-1}(x')} P(s'|s^{(2)},a) \qquad (\phi_1(s^{(1)}) = \phi_1(s^{(2)})) \\
&= P(y'|s^{(2)},a).
\end{aligned}
$$

On the second line, $\phi_1(\phi_{12}^{-1}(y'))$ is a set of abstract states in $\phi_1(\mathcal{S})$, formed by mapping each element of $\phi_{12}^{-1}(y')$ with ϕ_1 (recall that a set does not contain duplicate elements). The next step follows from the fact that any equivalence class in $\mathcal{S}$ induced by ϕ_{12} can always be partitioned into disjoint subsets, where each subset is a *complete* equivalence class under ϕ_1. Therefore, when we calculate $P(y'|s^{(1)},a)$, we can first sum over each smaller equivalence class under ϕ_1, and those probabilities will be the same for $s^{(1)}$ and $s^{(2)}$ as these two states are equivalent under ϕ_1 and the smaller equivalence classes are complete. As a consequence, the outer sum is also equal, and the result follows. □

2 Approximate abstractions

In practice, exact abstractions are hard to find and verify, so we want our theory to handle approximate abstractions.

Definition 2 (*lifting*)**.** For any function f that operates on $\phi(\mathcal{S})$, let $[f]_M$ denote its *lifted* version, which is a function over $\mathcal{S}$, defined as $[f]_M(s) := f(\phi(s))$. Similarly we can also lift a state-action value function. Lifting a real-valued function f over states can also be expressed in vector form: $[f]_M = \Phi^\top f$.

Definition 3 (Approximate abstractions)**.** Given MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and state abstraction ϕ that operates on $\mathcal{S}$, define the following types of abstractions:

1. ϕ is an $\epsilon_{\pi^\star}$-approximate $\pi^\star$-irrelevant abstraction, if there exists an abstract policy $\pi : \phi(\mathcal{S}) \to \mathcal{A}$, such that $\|V_M^\star - V_M^{[\pi]_M}\|_\infty \le \epsilon_{\pi^\star}$.
2. ϕ is an $\epsilon_{Q^\star}$-approximate $Q^\star$-irrelevant abstraction if there exists an abstract Q-value function $f : \phi(\mathcal{S}) \times \mathcal{A} \to \mathbb{R}$, such that $\|[f]_M - Q_M^\star\|_\infty \le \epsilon_{Q^\star}$.
3. ϕ is an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction if for any $s^{(1)}$ and $s^{(2)}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\forall a \in \mathcal{A}$,

$$|R(s^{(1)}, a) - R(s^{(2)}, a)| \le \epsilon_R, \quad \left\|\Phi P(s^{(1)}, a) - \Phi P(s^{(2)}, a)\right\|_1 \le \epsilon_P. \tag{3}$$

Note that Definition 1 is recovered when all approximation errors are set to 0.

The following theorem characterizes the relationship between the 3 types of approximate abstractions, with Theorem 1 as a direct corollary.

Theorem 2. *(1) If ϕ is an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction, then ϕ is also an approximate $Q^\star$-irrelevant abstraction with approximation error $\epsilon_{Q^\star} = \frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}$.*
(2) If ϕ is an $\epsilon_{Q^\star}$-approximate $Q^\star$-irrelevant abstraction, then ϕ is also an approximate $\pi^\star$-irrelevant abstraction with approximation error $\epsilon_{\pi^\star} = 2\epsilon_{Q^\star}/(1-\gamma)$.

A useful lemma for proving Theorem 2:

Lemma 3. *Let ϕ be an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction of M. Given any distributions $\{p_x : x \in \phi(\mathcal{S})\}$ where each p_x is supported on $\phi^{-1}(x)$, define $M_\phi = (\phi(\mathcal{S}), \mathcal{A}, P_\phi, R_\phi, \gamma)$, where $R_\phi(x, a) = \mathbb{E}_{s \sim p_x}[R(s, a)]$, and $P_\phi(x'|x, a) = \mathbb{E}_{s \sim p_x}[P(x'|s, a)]$. Then for any $s \in \mathcal{S}, a \in \mathcal{A}$,*

$$|R_\phi(\phi(s), a) - R(s, a)| \le \epsilon_R, \quad \|P_\phi(x, a) - \Phi P(s, a)\|_1 \le \epsilon_P.$$

Proof. We only prove for the transition part; the reward part follows from a similar (and easier) argument. Consider any fixed x and a. Let $q_s := \Phi P(s, a)$. By the definition of approximate bisimulation we have $\|q_{s^{(1)}} - q_{s^{(2)}}\|_1 \le \epsilon_P$ for any $\phi(s^{(1)}) = \phi(s^{(2)})$. The LHS of the claim on transition function is (let $x := \phi(s)$)

$$\begin{aligned}
&\|P_\phi(x, a) - \Phi P(s, a)\|_1 \\
=& \left\| \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s}) q_{\tilde{s}} - q_s \right\|_1 = \left\| \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s})(q_{\tilde{s}} - q_s) \right\|_1 \\
\le& \sum_{\tilde{s} \in \phi^{-1}(x)} \left\| p_x(\tilde{s})(q_{\tilde{s}} - q_s) \right\|_1 \le \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s}) \epsilon_P = \epsilon_P.
\end{aligned}$$

□

Proof of Theorem 2. Claim (2) follows directly from Lemma 4 in our first note, by using $\pi_{[f]_M}$ as the approximately optimal policy. It remains to prove Claim (1).

Define M_ϕ to be an abstract model as in Lemma 3 w.r.t. arbitrary distributions $\{p_x\}$. We will use $Q^\star_{M_\phi}$ as the f function in the definition of approximate $Q^\star$-irrelevance, and upper bound $\|[Q^\star_{M_\phi}]_M - Q^\star_M\|_\infty$ as:

$$\left\|[Q^\star_{M_\phi}]_M - Q^\star_M\right\|_\infty \leq \frac{1}{1-\gamma}\left\|[Q^\star_{M_\phi}]_M - \mathcal{T}[Q^\star_{M_\phi}]_M\right\|_\infty = \frac{1}{1-\gamma}\left\|[\mathcal{T}_{M_\phi}Q^\star_{M_\phi}]_M - \mathcal{T}[Q^\star_{M_\phi}]_M\right\|_\infty.$$

For any (s,a),

$$\begin{aligned}
&|([\mathcal{T}_{M_\phi}Q^\star_{M_\phi}]_M)(s,a) - (\mathcal{T}[Q^\star_{M_\phi}]_M)(s,a)| \\
&= |(\mathcal{T}_{M_\phi}Q^\star_{M_\phi})(\phi(s),a) - (\mathcal{T}[Q^\star_{M_\phi}]_M)(s,a)| \\
&= |R_\phi(\phi(s),a) + \gamma\langle P_\phi(\phi(s),a), V^\star_{M_\phi}\rangle - R(s,a) - \gamma\langle P(s,a), [V^\star_{M_\phi}]_M\rangle| \\
&\leq \epsilon_R + \gamma\left|\langle P_\phi(\phi(s),a), V^\star_{M_\phi}\rangle - \langle P(s,a), \Phi^\top V^\star_{M_\phi}\rangle\right| \\
&= \epsilon_R + \gamma\left|\langle P_\phi(\phi(s),a), V^\star_{M_\phi}\rangle - \langle \Phi P(s,a), V^\star_{M_\phi}\rangle\right| \qquad (*) \\
&\leq \epsilon_R + \gamma\epsilon_P\|V^\star_{M_\phi} - \tfrac{R_{\max}}{2(1-\gamma)}\mathbf{1}\|_\infty \\
&\leq \epsilon_R + \gamma\epsilon_P R_{\max}/(2(1-\gamma)).
\end{aligned}$$

In step (*), we notice that $[V^\star_{M_\phi}]_M$ is piece-wise constant, so when we take its dot-product with $P(s,a)$, we essentially first collapse $P(s,a)$ onto $\phi(\mathcal{S})$ (which is done by the Φ operator) and then take its dot-product with $V^\star_{M_\phi}$. The rest of the proof is similar to that of the simulation lemma. □

On monotonicity of approximation errors When we refine an abstraction ϕ, do we get better approximation? In fact, approximation errors are only monotone for $\pi^\star$- and $Q^\star$-irrelevance, but can be non-monotone for bisimulation. The former types of definitions are very much like *realizability* assumptions in supervised learning, and approximation errors monotonically decrease when function spaces get richer. For bisimulation, we will see later when we study Fitted Q-Iteration, that the criterion of bisimulation is essentially saying that the function space must be *closed* under Bellman update, and having more functions in the space can possibly break a closedness property.[2]

3 Bounding the loss of abstract models

The previous sections define different notions abstractions and their relationships. But what happens when we actually build a model using any type of abstractions and plan using the model? Is the output policy near-optimal? For this section we focus on approximation errors only, and will discuss estimation errors, that is, finite sample effects, in the next section.

3.1 ϕ is an approximate bisimulation

If we are given an (ϵ_R, ϵ_P)-approximate bisimulation abstraction and construct an abstract model M_ϕ as in Lemma 3, how lossy is $\pi^\star_{M_\phi}$? By applying both claims in Theorem 2 we obtain $\frac{2\epsilon_R}{(1-\gamma)^2} + \frac{\gamma\epsilon_P R_{\max}}{(1-\gamma)^3}$,

[2] My collaborators and I also referred to it as"completeness" in our recent paper [10].

which turns out to be loose. Here we provide a tighter analysis.

Theorem 4. *Let ϕ be an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction of M, and M_ϕ be an abstract model defined as in Lemma 3 with arbitrary distributions $\{p_x\}$, then*

$$\left\| V_M^\star - V_M^{[\pi^\star_{M_\phi}]_M} \right\|_\infty \leq \frac{2\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{(1-\gamma)^2}.$$

Proof. We first prove that for any *abstract* policy $\pi : \phi(\mathcal{S}) \to \mathcal{A}$,

$$\left\| [V^\pi_{M_\phi}]_M - V_M^{[\pi]_M} \right\|_\infty \leq \frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}. \tag{4}$$

To prove this, first recall the contraction property of policy-specific Bellman update operator for state-value functions, which implies that

$$\left\| [V^\pi_{M_\phi}]_M - V_M^{[\pi]_M} \right\|_\infty \leq \frac{1}{1-\gamma} \left\| [V^\pi_{M_\phi}]_M - \mathcal{T}^{[\pi]_M}[V^\pi_{M_\phi}]_M \right\|_\infty = \frac{1}{1-\gamma} \left\| [\mathcal{T}^\pi_{M_\phi} V^\pi_{M_\phi}]_M - \mathcal{T}^{[\pi]_M}[V^\pi_{M_\phi}]_M \right\|_\infty.$$

For notation simplicity let $R^{\pi'}(s) := R(s, \pi'(s))$ and $P^{\pi'}(s) := P(s, \pi'(s))$. For any $s \in \mathcal{S}$,

$$\begin{aligned}
&|[\mathcal{T}^\pi_{M_\phi} V^\pi_{M_\phi}]_M(s) - \mathcal{T}^{[\pi]_M}[V^\pi_{M_\phi}]_M(s)| \\
&= |(\mathcal{T}^\pi_{M_\phi} V^\pi_{M_\phi})(\phi(s)) - \mathcal{T}^{[\pi]_M}[V^\pi_{M_\phi}]_M(s)| \\
&= |R^\pi_\phi(\phi(s)) + \gamma \langle P^\pi_\phi(\phi(s)), V^\pi_{M_\phi} \rangle - R^{[\pi]_M}(s) - \gamma \langle P^{[\pi]_M}(s), V_M^{[\pi]_M} \rangle| \\
&\leq \epsilon_R + \gamma |\langle P^\pi_\phi(\phi(s)), V^\pi_{M_\phi} \rangle - \langle P^{[\pi]_M}(s), [V^\pi_{M_\phi}]_M \rangle| \\
&= \epsilon_R + \gamma \left| \langle P^\pi_\phi(\phi(s)), V^\pi_{M_\phi} \rangle - \langle \Phi P^{[\pi]_M}(s), V^\pi_{M_\phi} \rangle \right| \\
&\leq \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)}.
\end{aligned}$$

Now that we have a uniform upper bound on evaluation error, it might be attempting to argue that we under-estimate $\pi^\star_M$ and over-estimate $\pi^\star_{M_\phi}$ at most this much, hence the decision loss is twice the evaluation error. This argument does not apply here because $\pi^\star_M$ cannot be necessarily expressed as a lifted abstract policy when ϕ is not an exact bisimulation!

Instead we can use the following argument: for any $s \in \mathcal{S}$,

$$\begin{aligned}
V^\star_M(s) - V_M^{[\pi^\star_{M_\phi}]_M}(s) &= V^\star_M(s) - V^\star_{M_\phi}(\phi(s)) + V^\star_{M_\phi}(\phi(s)) - V_M^{[\pi^\star_{M_\phi}]_M}(s) \\
&\leq \left\| Q^\star_M - [Q^\star_{M_\phi}]_M \right\|_\infty + \left\| [V_{M_\phi}^{\pi^\star_{M_\phi}}]_M - V_M^{[\pi^\star_{M_\phi}]_M} \right\|_\infty.
\end{aligned}$$

Here both terms can be bounded by $\frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}$ but for different reasons: the bound applies to the first term due to Claim (1) of Theorem 2, and applies to the second term through Eq.(4) as $\pi^\star_{M_\phi}$ is an abstract policy. □

3.2 ϕ is approximately $Q^\star$-irrelevant

When ϕ is an approximate $Q^\star$-irrelevant abstraction with low approximation error, building a model based on ϕ may not seem a good idea, as the transitions and rewards for states with similiar $Q^\star$-values may be drastically different, and the parameters of M_ϕ (as in Lemma 3) may not be meaningful at all.

Perhaps surprisingly, we can show that M_ϕ produces a near-optimal $Q^\star$-function hence a near-optimal policy.[3]

Theorem 5. *Let ϕ be an $\epsilon_{Q^\star}$-approximate $Q^\star$-irrelevant abstraction for M. Then, for M_ϕ constructed as in Lemma 3 with arbitrary distributions $\{p_x\}$, we have $\|[Q^\star_{M_\phi}]_M - Q^\star_M\|_\infty \le 2\epsilon_{Q^\star}/(1-\gamma)$.*

Exact $Q^\star$-irrelevance To develop intuition, let's see what happens when ϕ is an exact $Q^\star$-irrelevant abstraction: we can prove that $[Q^\star_{M_\phi}]_M = Q^\star_M$, despite that the dynamics and rewards in M_ϕ "do not make sense". In particular, we know that for any $s^{(1)}$ and $s^{(2)}$ aggregated by ϕ, for any $a \in \mathcal{A}$,

$$R(s^{(1)},a) + \gamma\langle P(s^{(1)},a), V^\star_M\rangle = Q^\star(s^{(1)},a) = Q^\star(s^{(2)},a) = R(s^{(2)},a) + \gamma\langle P(s^{(2)},a), V^\star_M\rangle.$$

This equation tells us that, although ϕ aggregates states that can have very different rewards and dynamics, they at least share one thing: the Bellman operator updates $Q^\star_M$ in exactly the same way at $s^{(1)}$ and $s^{(2)}$ (for any action).

Let $[Q^\star_M]_\phi(x,a) = Q^\star_M(s,a)$ for any $s \in \phi^{-1}(x)$; note that the notation $[\cdot]_\phi$ can only be applied to functions that are piece-wise constant under ϕ. We now show that $[Q^\star_M]_\phi$ is the fixed point of $\mathcal{T}_{M_\phi}$, which proves the claim. This is because, for any $x \in \phi(\mathcal{S}), a \in \mathcal{A}$, let s be any state in $\phi^{-1}(x)$:

$$\begin{aligned}
(\mathcal{T}_{M_\phi}[Q^\star_M]_\phi)(x,a) &= R_\phi(x,a) + \gamma\langle P_\phi(x,a), [V^\star_M]_\phi\rangle \\
&= \sum_{s\in\phi^{-1}(x)} p_x(s)\left(R(s,a) + \gamma\langle \Phi P(s,a), [V^\star_M]_\phi\rangle\right) \\
&= \sum_{s\in\phi^{-1}(x)} p_x(s)\left(R(s,a) + \gamma\langle P(s,a), V^\star_M\rangle\right) \\
&= \sum_{s\in\phi^{-1}(x)} p_x(s)\,[Q^\star_M]_\phi(x,a) = [Q^\star_M]_\phi(x,a).
\end{aligned}$$

The approximate case The more general case is much trickier, as $Q^\star_M$ is not piece-wise constant when ϕ is not exactly $Q^\star$-irrelevant, so we cannot apply $\mathcal{T}_{M_\phi}$ to it.

To get around this issue, define a new MDP $M'_\phi = (\mathcal{S}, \mathcal{A}, P'_\phi, R'_\phi, \gamma)$, with

$$R'_\phi(s,a) = \mathbb{E}_{\tilde{s}\sim p_{\phi(s)}}[R(\tilde{s},a)], \qquad P'_\phi(s'|s,a) = \mathbb{E}_{\tilde{s}\sim p_{\phi(s)}}[P(s'|\tilde{s},a)].$$

Recall that $\{p_x\}$ are a set of arbitrary distributions and we use them as weights for defining M_ϕ. The model here, $M_{\phi'}$, also combines parameters from aggregated states, but is defined over the *primitive* state space. This seemingly crazy model has two important properties: (1) Its optimal Q-value function coincides with that of M_ϕ (after lifting), and (2) It's defined over $\mathcal{S}$ so we can apply its Bellman operator to $Q^\star_M$.

[3] In fact, this is why the guarantee of Delayed Q-learning, a PAC-MDP algorithm, can be extended to $Q^\star$-irrelevant abstractions; see Section 8.2.3 of Lihong Li's thesis [5].

We first prove that $[Q^\star_{M_\phi}]_M = Q^\star_{M'_\phi}$, by showing that $\mathcal{T}_{M'_\phi}[Q^\star_{M_\phi}]_M = [Q^\star_{M_\phi}]_M$:

$$\begin{aligned}
(\mathcal{T}_{M'_\phi}[Q^\star_{M_\phi}]_M)(s,a) &= R'_\phi(s,a) + \gamma\langle P'_\phi(s,a), [V^\star_{M_\phi}]_M\rangle \\
&= \sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) \left(R(\tilde{s},a) + \gamma\langle P(\tilde{s},a), [V^\star_{M_\phi}]_M\right) \\
&= \sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) R(\tilde{s},a) + \sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s})\gamma\langle \Phi P(\tilde{s},a), V^\star_{M_\phi}\rangle \\
&= R_\phi(\phi(s),a) + \gamma\langle P_\phi(\phi(s),a), V^\star_{M_\phi}\rangle \\
&= Q^\star_{M_\phi}(\phi(s),a) = [Q^\star_{M_\phi}]_M(s,a).
\end{aligned}$$

With this result, we have

$$\left\|[Q^\star_{M_\phi}]_M - Q^\star_M\right\|_\infty = \left\|Q^\star_{M'_\phi} - Q^\star_M\right\|_\infty \le \frac{1}{1-\gamma}\left\|\mathcal{T}_{M'_\phi}Q^\star_M - Q^\star_M\right\|_\infty.$$

And

$$\begin{aligned}
&|(\mathcal{T}_{M'_\phi}Q^\star_M)(s,a) - Q^\star_M(s,a)| \\
&= |R'_\phi(s,a) + \gamma\langle P'_\phi(s,a), V^\star_M\rangle - Q^\star_M(s,a)| \\
&= \left|\left(\sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s})\left(R(\tilde{s},a) + \gamma\langle P(\tilde{s},a), V^\star_M\rangle\right)\right) - Q^\star_M(s,a)\right| \\
&= \left|\sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s})\left(Q^\star_M(\tilde{s},a) - Q^\star_M(s,a)\right)\right| \\
&\le \left|\sum_{\tilde{s}:\phi(\tilde{s})=\phi(s)} p_x(\tilde{s})(2\epsilon_{Q^\star})\right| = 2\epsilon_{Q^\star}.
\end{aligned}$$

4 Finite sample analysis

We briefly discuss the finite sample guarantees of certainty-equivalence RL after pre-processing data using a state abstraction ϕ [8].

As before, we assume that the dataset, $D = \{D_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, is formed by sampling rewards and transitions from each (s,a) a number of times. Previously we made the simplification assumption that $|D_{s,a}|$ is a constant for all (s,a); here we remove this assumption and allow their sizes to vary, for the following reason: when the primitive state space $\mathcal{S}$ is very large and the amount of total data is limited, there can be many states where we don't even have any data, so assuming constant $|D_{s,a}|$ (which is at least 1) is unrealistic in this scenario. In fact, such a scenario is exactly where abstractions can be very useful due to their generalization effects.

In particular, the effective sample size that will enter our analysis is

$$n_\phi(D) := \min_{x\in\phi(\mathcal{S}),a\in\mathcal{A}} |D_{x,a}|, \quad \text{where} \quad D_{x,a} := \sum_{s\in\phi^{-1}(x)} |D_{s,a}|.$$

In words, $n_\phi(D)$ is the least number of samples for any *abstract* state-action pair. Note that even if $|D_{s,a}| = 0$ for many (s,a), if ϕ aggregate states aggressively and data is relatively uniform over all

abstract states, we may still have a reasonably large $n_\phi(D)$. Our loss bound will depend on $n_\phi(D)$ and the approximation error of the representation ϕ, but will not incur any dependence on the sample size of individual states (which implicitly depends on $|\mathcal{S}|$).

Recall that in the note on tabular RL we studied two approaches to the finite sample analyses of certainty-equivalence: one through $\max_\pi \|V^\pi_M - V^\pi_{\widehat{M}}\|_\infty$ (uniform bound of policy evaluation errors) and the other through $\|Q^\star_M - Q^\star_{\widehat{M}}\|$. To extend the first approach to the setting of abstractions we need to assume approximate bisimulation, and to extend the second we only need approximate $Q^\star$-irrelevance. We discuss the second approach in details below, which covers some important desiderata that also applies to the extension of the first approach.

Before that, we need a few more notations: Let $\widehat{M}_\phi = (\phi(\mathcal{S}), \mathcal{A}, \widehat{P}_\phi, \widehat{R}_\phi, \gamma)$ be the estimated model using the abstract representation. Let $M_\phi = (\phi(\mathcal{S}), \mathcal{A}, P_\phi, R_\phi, \gamma)$ be the following MDP:

$$R_\phi(x,a) = \frac{\sum_{\tilde{s}\in\phi^{-1}(x)} |D_{\tilde{s},a}| R(s,a)}{|D_{\phi(s),a}|}, \quad P_\phi(x'|x,a) = \frac{\sum_{\tilde{s}\in\phi^{-1}(x)} |D_{\tilde{s},a}| P(x'|s,a)}{|D_{\phi(s),a}|}.$$

This is essentially the definition of M_ϕ in Lemma 3 with $p_x(s) \propto |D_{s,a}|$. In words, M_ϕ is the "expectation" of $\widehat{M}_\phi$ w.r.t. the randomness in data. If $|D_{s,a}|$ gets multiplied by the same constant for all (s,a) and goes to infinity, M_ϕ is what $\widehat{M}_\phi$ converges to in the limit.

Bounding $\left\|Q^\star_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty$: We bound it by introducing an intermediate term:

$$\left\|Q^\star_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty \le \left\|Q^\star_M - [Q^\star_{M_\phi}]_M\right\|_\infty + \left\|[Q^\star_{M_\phi}]_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty.$$

We have already bounded the first term on the RHS in Theorem 2 for approximate bisimulations and Theorem 5 for approximate $Q^\star$-irrelevant abstractions, respectively, so it remains to deal with the second term. A few comments on this decomposition before we dive into the analysis:

1. Using terminologies from statistical learning theory, we will call the first term *approximation error*, and the second term *estimation error*.

2. Approximation error reflects the fidelity of a representation ϕ to the true model M. It does not vanish with more data. In fact, this is what we have to pay in the limit of infinite data. The finer ϕ is, the better approximation we get (see more detailed discussions at the end of Section 2).

3. Estimation error reflects how fast the estimated model converges to its "expected" version. It goes to 0 as sample size goes to infinity, and it has nothing to do with whether ϕ is a legit representation for M. The coarser ϕ, the lower estimation error we get.

4. So here is the trade-off: to guarantee low approximation error we want a fine ϕ, but to guarantee low estimation error we want a coarse ϕ. In general, the best balance depends on the sample size [8].

Now let's bound $\left\|[Q^\star_{M_\phi}]_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty$, $\widehat{M}_\phi$ converges to M_ϕ in the limit, so the second term should go to 0 as $n_\phi(D)$ goes to infinity, and the fact that ϕ is an inexact abstraction for M is irrelevant here. However, we cannot argue that data in $D_{x,a}$ can be viewed as if they were sampled from $P_\phi(x,a)$, since the subsets of data from different s have independent but non-identical distributions.

Fortunately, Hoeffding's inequality applies to independent but non-identical distributions, and we can leverage this property to get around the issue:

$$\begin{aligned}&\left\|[Q^\star_{M_\phi}]_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty = \left\|Q^\star_{M_\phi} - Q^\star_{\widehat{M}_\phi}\right\|_\infty \\ \leq \frac{1}{1-\gamma}&\left\|Q^\star_{M_\phi} - \mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi}\right\|_\infty = \frac{1}{1-\gamma}\left\|\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi} - \mathcal{T}_{M_\phi} Q^\star_{M_\phi}\right\|_\infty.\end{aligned}$$

For each $(x, a) \in \phi(S) \times \mathcal{A}$,

$$\begin{aligned}&|(\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi})(x,a) - (\mathcal{T}_{M_\phi} Q^\star_{M_\phi})(x,a)| \\ &= |\widehat{R}_\phi(x,a) + \gamma\langle \widehat{P}_\phi(x,a), V^\star_{M_\phi}\rangle - R_\phi(x,a) - \gamma\langle P_\phi(x,a), V^\star_{M_\phi}\rangle| \\ &= \left|\frac{1}{|D_{x,a}|}\sum_{s\in\phi^{-1}(x)}\sum_{(r,s')\in D_{s,a}}\left(r + \gamma V^\star_{M_\phi}(\phi(s')) - R(s,a) - \gamma\langle P(s,a), [V^\star_{M_\phi}]_M\rangle\right)\right|.\end{aligned}$$

If we view the nested sum as a flat sum, the expression is the sum of the differences between random variables $r + \gamma V^\star_{M_\phi}(s')$ and their expectation w.r.t. the randomness of (r, s'), so Hoeffding's inequality applies (although for different $s \in \phi^{-1}(x)$ the random variables have non-identical distributions): with probability at least $1 - \delta$,

$$\left\|\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi} - \mathcal{T}_{M_\phi} Q^\star_{M_\phi}\right\|_\infty \leq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n_\phi(D)}\ln\frac{2|\phi(\mathcal{S})\times\mathcal{A}|}{\delta}}.$$

This completes the analysis.

References

[1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[2] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

[3] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1):163–223, 2003.

[4] Balaraman Ravindran. *An algebraic approach to abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2004.

[5] Lihong Li. *A unifying framework for computational reinforcement learning theory*. PhD thesis, Rutgers, The State University of New Jersey, 2009.

[6] Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 752–757, 2005.

[7] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[8] Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction Selection in Model-based Reinforcement Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 179–188, 2015.

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[10] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. On oracle-efficient PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems 31 (to appear)*, 2018.