Extreme State Aggregation Beyond MDPs

Marcus Hutter

Research School of Computer Science
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

12 July 2014

Abstract

We consider a Reinforcement Learning setup where an agent interacts with an environment in observation-reward-action cycles without any (esp. MDP) assumptions on the environment. State aggregation and more generally feature reinforcement learning is concerned with mapping histories/raw-states to reduced/aggregated states. The idea behind both is that the resulting reduced process (approximately) forms a small stationary finite-state MDP, which can then be efficiently solved or learnt. We considerably generalize existing aggregation results by showing that even if the reduced process is not an MDP, the (q-)value functions and (optimal) policies of an associated MDP with same state-space size solve the original problem, as long as the solution can approximately be represented as a function of the reduced states. This implies an upper bound on the required state space size that holds uniformly for all RL problems. It may also explain why RL algorithms designed for MDPs sometimes perform well beyond MDPs.

Contents

Keywords

state aggregation, reinforcement learning, non-MDP.

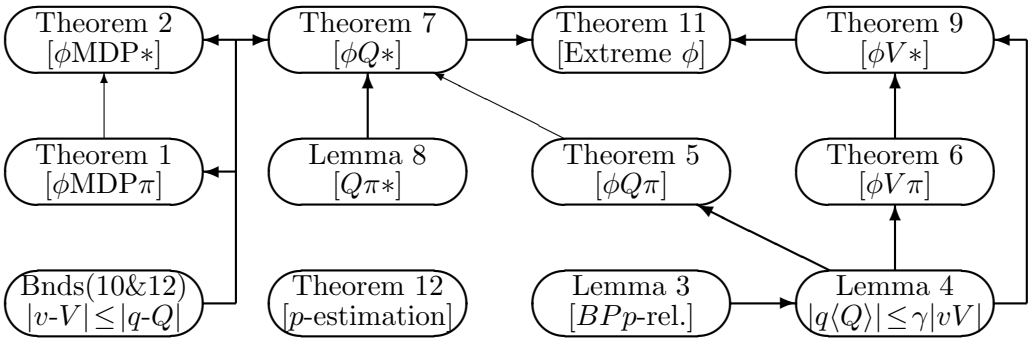
1 Introduction

In *Reinforcement Learning* (RL) [SB98], an *agent* Π takes actions in some *environment* P and observes its consequences and is rewarded for them. A well-understood and efficiently solvable [Put94] and efficiently learnable [SLL09, LH12] case is where the environment is (modelled as) a finite-state stationary *Markov Decision Process* (MDP). Unfortunately most interesting real-world problems P are neither finite-state, nor stationary, nor Markov. One way of dealing with this mismatch is to somehow transform the real-world problem into a small MDP: *Feature Reinforcement Learning* (FRL) [Hut09c] and U-tree [McC96] deal with the case of arbitrary unknown environments, while state aggregation assumes the environment is a large known stationary MDP [GDG03, FPP04]. The former maps histories into states (Section 2), the latter groups raw states into aggregated states.

Here we follow the FRL approach and terminology, since it is arguably most general: It subsumes the cases where the original process P is an MDP, a k-order MDP, a POMDP, and others (Section 3). Thinking in terms of histories also naturally stifles any temptation of a naive frequency estimate of P (no history ever repeats). Finally we find the history vs state terminologically somewhat neater than raw state vs aggregated state.

More importantly, we consider maps ϕ from histories to states for which the reduced process P_ϕ is not (even approximately) an MDP (Section 4). At first this seems to defeat the original purpose, namely of reducing P to a well-understood and efficiently solvable problem class, namely small MDPs. The main novel contribution of this paper is to show that there is still an associated finite-state stationary MDP p whose solution (approximately) solves the original problem P, as long as the solution can still be represented (Section 5). Indeed, we provide an upper bound on the required state space size that holds uniformly for all P (Section 6). While these are interesting theoretical insights, it is a-priori not clear whether they could by utilized to design (better) RL algorithms. We also show how to learn p from experience (Section 7), and sketch an overall learning algorithm and regret/PAC analysis based on our main theorems (Section 8). We briefly discuss how to relax one of the conditions in our main theorems by permuting actions (Section 9). We conclude with an outlook on future work and open problems (Section 10). A list of notation can be found in Appendix A.

The diagram below depicts the dependencies between our results:



2 Feature Markov Decision Processes (ΦMDP)

This section formally describes the setup of [Hut09c]. It consists of the agent-environment framework and maps ϕ from observation-reward-action histories to MDP states. This arrangement is called "Feature MDP" or short ΦMDP. We use upper-case letters P, Q, V, and Π for the Probability, (Q-)Value, and Policy of the original (agent-environment interactive) Process, and lower-case letters p, q, v, and π for the probability, (q-)value, and policy of the (reduced/aggregated) MDP.

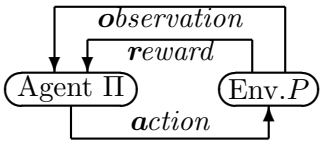
Agent-environment setup [Hut09c]. We start with the standard agent-environment setup [RN10] in which an agent Π interacts with an environment P. The agent can choose from actions $a \in \mathcal{A}$ and the environment provides observations $o \in \mathcal{O}$ and real-valued rewards $r \in \mathcal{R} \subseteq [0;1]$ to the agent. This happens in cycles $t = 1,2,3,...$: At time t, after observing o_t and receiving reward r_t, the agent takes action a_t based on history

$$h_t \;:=\; o_1 r_1 a_1 ... o_{t-1} r_{t-1} a_{t-1} o_t r_t \;\in\; \mathcal{H}_t := (\mathcal{O} \times \mathcal{R} \times \mathcal{A})^{t-1} \times \mathcal{O} \times \mathcal{R}$$

Then the next cycle $t+1$ starts. The agent's objective is to maximize its long-term reward. To avoid integrals and densities, we assume spaces $\mathcal{O}$ and $\mathcal{R}$ are finite. They may be huge, so this is not really restrictive. Indeed, the ΦMDP framework has been specifically developed for huge observation spaces. Generalization to continuous $\mathcal{O}$ and $\mathcal{R}$ is routine [Hut09a]. Furthermore we assume that $\mathcal{A}$ is finite and smallish, which is restrictive. Potential extensions to continuous $\mathcal{A}$ are discussed in Section 10.

The agent and environment may be viewed as a pair of interlocking functions of the history $\mathcal{H} := (\mathcal{O} \times \mathcal{R} \times \mathcal{A})^* \times \mathcal{O} \times \mathcal{R}$:

$$\begin{aligned} &\text{Env. } P : \mathcal{H} \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}, \qquad P(o_{t+1} r_{t+1} | h_t a_t), \\ &\text{Agent } \Pi : \mathcal{H} \rightsquigarrow \mathcal{A}, \qquad\quad \Pi(a_t | h_t) \quad \text{or} \quad a_t = \Pi(h_t), \end{aligned}$$



where $\rightsquigarrow$ indicates that mappings $\rightarrow$ are in general stochastic. We make no (stationarity or Markov or other) assumption on environment P. For most parts, environment P is assumed to be fixed, so dependencies on P will be suppressed. For convenience and since optimal policies can be chosen to be deterministic, we consider deterministic policies $a_t = \Pi(h_t)$ only.

Value functions, optimal Policies, and history Bellman equations. We measure the performance of a policy Π in terms of the P-expected γ-discounted reward sum ($0\leq\gamma<1$), called (Q-)Value of Policy Π at history h_t (and action a_t)

$$V^\Pi(h_t) := \mathbb{E}^\Pi[R_{t+1}|h_t] \quad \text{and} \quad Q^\Pi(h_t,a_t) := \mathbb{E}^\Pi[R_{t+1}|h_t a_t], \quad \text{where} \quad R_t := \sum_{\tau=t}^{\infty}\gamma^{\tau-t}r_\tau$$

The optimal Policy and (Q-)Value functions are

$$\begin{aligned} V^*(h_t) &:= \max_\Pi V^\Pi(h_t) \quad \text{and} \quad Q^*(h_t,a_t) := \max_\Pi Q^\Pi(h_t,a_t), \\ &\text{where} \quad \Pi^* :\in \arg\max_\Pi V^\Pi(\epsilon) \end{aligned} \tag{1}$$

The maximum over all policies Π always exists [LH14] but may not be unique, in which case argmax denotes the set of optimal policies and Π^* denotes a representative or the whole set of optimal policies. Despite being history-based we can write down (pseudo)recursive Bellman (optimality) equations for the (optimal) (Q-)Values [Hut05, Sec.4.2]:

$$Q^\Pi(h_t,a_t) = \sum_{o_{t+1}r_{t+1}} P(o_{t+1}r_{t+1}|h_t a_t)[r_{t+1}+\gamma V^\Pi(h_{t+1})], \; V^\Pi(h_t)=Q^\Pi(h_t,\Pi(h_t)) \tag{2}$$

$$Q^*(h_t,a_t) = \sum_{o_{t+1}r_{t+1}} P(o_{t+1}r_{t+1}|h_t a_t)[r_{t+1}+\gamma V^*(h_{t+1})], \; V^*(h_t)=\max_{a_t\in\mathcal{A}} Q^*(h_t,a_t) \tag{3}$$

$$\Pi^*(h_t) \in \arg\max_{a_t\in\mathcal{A}} Q^*(h_t,a_t) \tag{4}$$

Unlike their classical state-space cousins (see below), they are *not* self-consistency equations: The r.h.s. refers to a longer history h_{t+1} which is always different from the history h_t on the l.h.s, which precludes any learning algorithm based on estimating the frequency of state/history visits. Still the recursions will be convenient for the mathematical development.

From histories to states (ϕ). The space of histories is huge and unwieldy and no history ever repeats. Standard ways of dealing with this are to define a similarity metric on histories [McC96] or to aggregate histories [Hut09c]. We pursue the latter via a feature map $\phi:\mathcal{H}\to\mathcal{S}$ which reduces histories $h_t\in\mathcal{H}$ to states $s_t:=\phi(h_t)\in\mathcal{S}$. W.l.g. we assume that ϕ is surjective. We also assume that state space $\mathcal{S}$ is finite; indeed we are interested in small $\mathcal{S}$. This corresponds and indeed is equivalent to a partitioning of histories $\{\phi^{-1}(s):s\in\mathcal{S}\}$. Classical state aggregation usually uses the partitioning view [GDG03, Ort07], but the map notation is a bit more convenient here.

The state s_t is supposed to summarize all relevant information in history h_t, which lower bounds the size of $\mathcal{S}$. We pass from the complete history $o_1r_1a_1...o_nr_n$ to a 'reduced' history $s_1r_1a_1...s_nr_n$. Traditionally, 'relevant' means that the future is predictable from s_t (and a_t) alone, or technically that the reduced history forms a Markov decision process. This is precisely the condition this paper intends to lift (later).

From histories to MDPs. The probability of the successor states and rewards can be obtained by marginalization

$$P_\phi(s_{t+1}r_{t+1}|h_t a_t) \;:=\; \sum_{\tilde{o}_{t+1}:\phi(h_t a_t \tilde{o}_{t+1} r_{t+1})=s_{t+1}} P(\tilde{o}_{t+1}r_{t+1}|h_t a_t) \tag{5}$$

The reduced process P_ϕ is a Markov Decision Process, or Markov for short, if P_ϕ only depends on h_t through s_t, i.e. is the same for all histories mapped to the same state. Formally

$$P_\phi \in \text{MDP} \quad :\Longleftrightarrow \quad \exists p : P_\phi(s_{t+1}r_{t+1}|\tilde{h}_t a_t) = p(s_{t+1}r_{t+1}|s_t a_t) \;\; \forall \phi(\tilde{h}_t) = s_t \tag{6}$$

Here and elsewhere a quantifier such as $\forall \phi(\tilde{h}_t) = s_t$ shall mean: for all values of all involved variables consistent with the constraint $\phi(\tilde{h}_t) = s_t$. The MDP P_ϕ is assumed to be stationary, i.e. independent of t; another condition to be lifted later. Condition (6) is essentially the stochastic bisimulation condition generalized to histories and being somewhat more restrictive regarding rewards [GDG03]: It is a condition on the reward distribution, while [GDG03] constrains its expectation only. This could easily be rectified but is besides the point of this paper. The bisimulation metric [FPP04] is an approximate version of (6), which measures the deviation of P_ϕ from being an MDP.

Many problems P can be reduced (approximately) to stationary MDPs [Hut09c]: Full-information *games* such as chess with static opponent are already Markov, classical *physics* is approximately 2nd-order Markov, (conditional) i.i.d. processes such as *Bandits* have counting sufficient statistics, and for a *POMDP planning* problem, the belief vector is Markov.

Markov decision processes (MDP). We have used and continue to use upper-case letters V, Q, Π for the general process P. We will use lower-case letters v, q, π for (stationary) MDPs p. We use s and a for the current state and action, and s' and r' for successor state and reward. Consider a stationary finite-state MDP $p : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S} \times \mathcal{R}$ and stationary deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$. Only in Section 3 will this p be given by (6), but *in general p will be different from* (6). In any case, the p-expected γ-discounted reward sum, called (q-)value of (optimal) policy $\pi^{(*)}$ in MDP p, are given by the Bellman (optimality) equations

$$q^\pi(s,a) \;=\; \sum_{s'r'} p(s'r'|sa)[r' + \gamma v^\pi(s')] \quad \text{and} \quad v^\pi(s) = q^\pi(s,\pi(s)) \tag{7}$$

$$q^*(s,a) \;=\; \sum_{s'r'} p(s'r'|sa)[r' + \gamma v^*(s')] \quad \text{and} \quad v^*(s) = \max_a q^*(s,a) \tag{8}$$

$$\pi^*(s) \;\in\; \arg\max_a q^*(s,a). \quad \text{Note: } v^\pi(s) \le v^*(s),\; q^\pi(s,a) \le q^*(s,a) \tag{9}$$

Using $p(s'r'|sa) = p(r'|sas')p(s'|sa)$ we could also rewrite them in terms of transition matrix $p(s'|sa)$ and expected reward $\mathbb{E}[r'|sa]$ [SB98].

More notation. While our equations often assume or imply $s=s_t$, $a=a_t$, $s'=s_{t+1}$, $r'=r_{t+1}$, (and $h_{t+1}=h_tao'r'$) for some t, technically s,a,s',r' are *different* variables from all variables in history $h_n=o_1r_1a_1...o_tr_ta_to_{t+1}r_{t+1}...a_nr_n$. Less prone to confusion are $o=o_t$, $o'=o_{t+1}$, $h=h_t$, $h'=hao'r'$.

We call a function $f(h)$, piecewise constant or ϕ-uniform iff $f(h)=f(\tilde{h})$ for all $\phi(h)=\phi(\tilde{h})$. Here and elsewhere $\forall\phi(h)=\phi(\tilde{h})$ is short for $\forall h,\tilde{h}:\phi(h)=\phi(\tilde{h})$. Similarly $\forall s=\phi(h)$ is short for $\forall s,h:s=\phi(h)$. Etc.

The Iverson bracket, $[\![R]\!]:=1$ if R=true and $[\![R]\!]:=0$ if R=false, denotes the indicator function. Throughout, $\varepsilon,\delta\geq 0$ denote approximation accuracy. Note that this includes the exact $=0$ case.

We now show that if P reduces via ϕ to an MDP p, the solution of these equations yields (Q-)Values and optimal Policy of the original process P. This is not surprising and just a history-based versions of classical state-aggregation results [GDG03]. We state and prove them here, since notation and setup are somewhat different, and proof ideas and fragments will be reused later.

3 Exact Aggregation for $P_\phi\in$MDP

The following two theorems show that if ϕ reduces P to a stationary MDP via (5) and (6), then V and Q (and Π^*) essentially coincide with v and q (and π^*), where policy Π (Π^*) has to be assumed (will be shown) constant within each partition $\phi^{-1}(s)$. This allows to efficiently solve for (and learn in the case of unknown P) V and Q (and Π^*) in time polynomial in $\mathcal{S}$ by solving/learning (7) (or (8) and (9)) instead of (2) (or (3) and (4)).

Theorem 1 (ϕMDPπ) *Let ϕ be a reduction such that $P_\phi\in$MDP reduces to MDP p defined in (6), and let Π be some policy such that $\Pi(h)=\Pi(\tilde{h})$ for all $\phi(h)=\phi(\tilde{h})$. Then for all a and h it holds:*

$$V^\Pi(h)=v^\pi(s)\quad\textit{and}\quad Q^\Pi(h,a)=q^\pi(s,a),\quad\textit{where}\quad\pi(s):=\Pi(h)\quad\textit{and}\quad s=\phi(h)$$

Note that $\pi(s)$ is well-defined, since ϕ is surjective and $\Pi(h)$ is the same for all $h\in\phi^{-1}(s)$. The standard proof considers an m-horizon truncated MDP and induction on m and $m\to\infty$. Besides the adaptation to histories, the proof below is a slight variation that avoids such truncation and limit. This style will be useful later. We explain all steps in detail here, since variations will be utilize later.

Proof. Let $\delta:=\sup_{s=\phi(h),a}|q^\pi(s,a)-Q^\Pi(h,a)|$. Using $a':=\pi(s')=\Pi(h')$ for $s'=\phi(h')$ and (2) and (7) lets us bound the value difference

$$|v^\pi(s')-V^\Pi(h')|\;=\;|q^\pi(s',a')-Q^\Pi(h',a')|\;\leq\;\delta\quad\forall s'=\phi(h')\tag{10}$$

For any a and h, this implies

$$\begin{aligned}
Q^{\Pi}(h,a) &\overset{(a)}{=} \sum_{o'r'} P(o'r'|ha)[r'+\gamma V^{\Pi}(h')] && [h'=hao'r'] \\
&\overset{(b)}{\leqslant} \sum_{s'r'} \sum_{o':\phi(h')=s'} P(o'r'|ha)[r'+\gamma(v^{\pi}(s')\pm\delta)] && (11)\\
&\overset{(c)}{=} \sum_{s'r'} P_{\phi}(s'r'|ha)[r'+\gamma v^{\pi}(s')]\pm\gamma\delta \\
&\overset{(d)}{=} \sum_{s'r'} p(s'r'|sa)[r'+\gamma v^{\pi}(s')]\pm\gamma\delta && [s:=\phi(h)] \\
&\overset{(e)}{=} q^{\pi}(s,a)\pm\gamma\delta
\end{aligned}$$

(a) is just (2). In (b) we sum over all o' by first summing over all o' such that $\phi(hao'r')=s'$ and then summing over all s'. We have also upper/lower bounded $V^{\Pi}(h')$ via (10). (c) is the definition (5) of P_{ϕ} and pulls out $\gamma\delta$ using that probability P_{ϕ} sums to 1. (d) is the definition (6) of p. (e) is simply (7). The chain (11a-e) holds for all $s=\phi(h)$ and a, hence

$$\delta \;=\; \sup_{s=\phi(h),a} |q^{\pi}(s,a)-Q^{\Pi}(h,a)| \;\leq\; \gamma\delta \quad\Rightarrow\quad \delta\leq 0$$

Hence $v^{\pi}(s)=V^{\Pi}(h)$ and $q^{\pi}(s,a)=Q^{\Pi}(h,a)$ for all $s=\phi(h)$ and a. ■

Theorem 2 (ϕMDP$*$) *Let ϕ be a reduction such that $P_{\phi}\in$ MDP reduces to MDP p defined in* (6)*, Then for all a and h it holds:*

$$\Pi^*(h)=\pi^*(s) \quad \textit{and} \quad V^*(h)=v^*(s) \quad \textit{and} \quad Q^*(h,a)=q^*(s,a), \quad \textit{where} \quad s=\phi(h)$$

The core of the proof follows the same steps (11a-e) as for the previous theorem, but the rest is slightly different. Additionally we have to show that Π^* is piecewise constant (in Theorem 1 we assumed Π was).

Proof. Let $\delta:=\sup_{s=\phi(h),a}|q^*(s,a)-Q^*(h,a)|$. We can bound the value difference

$$|v^*(s)-V^*(h)| \overset{(a)}{=} |\max_a q^*(s,a)-\max_a Q^*(h,a)| \overset{(b)}{\leq} \max_a |q^*(s,a)-Q^*(h,a)| \overset{(c)}{\leq} \delta\;\forall s=\phi(h) \tag{12}$$

(a) follows from the definitions (3) and (8). (b) follows from the following general elementary frequently used bound

$$|\max_x f(x)-\max_x g(x)| \;\leq\; \max_x |f(x)-g(x)| \tag{13}$$

(c) follows from the definition of δ.

One now can show that $Q^*(h,a) \lessgtr q^*(s,a) \pm \gamma\delta$ for $s=\phi(h)$ by following exactly the same steps as (10a-e) just with Π and π replaced by $*$ and using (12) instead of (10), and using the Bellman optimality equations (3) and (8) instead of the Bellman equations (2) and (7). Also as before, this implies $\delta \leq \gamma\delta$, hence $\delta \leq 0$, hence $v^*(s)=V^*(h)$ and $q^*(s,a)=Q^*(h,a)$ for all $s=\phi(h)$ and a. Finally, the latter implies $\pi^*(s)=\text{argmax}_a q^*(s,a)=\text{argmax}_a Q^*(h,a)=\Pi^*(h)$. ■

Approximate aggregation results if P_ϕ is approximately MDP can also be derived [FPP04]. The core results in the next section show that aggregation is possible far beyond P_ϕ being approximately MDP.

4 Approximate Aggregation for General P

This section prepares for the main technical contribution of the paper in the next section. The key quantity to relate original and reduced Bellman equations is a form of stochastic inverse of ϕ, whose choice and analysis will be deferred to Section 7.

Dispersion probability B. Let $B_\phi : \mathcal{S}\times\mathcal{A} \rightsquigarrow \mathcal{H}$ be a probability distribution on finite histories for each state-action pair such that $B_\phi(h|sa)=0$ if $s\neq\phi(h)$. $B\equiv B_\phi$ may be viewed as a stochastic inverse of ϕ that assigns non-zero probability only to $h\in\phi^{-1}(s)$. The formal constraints we pose on B are

$$B(h|sa) \geq 0 \quad \text{and} \quad \sum_{h\in\mathcal{H}} B(h|sa) = \sum_{h:\phi(h)=s} B(h|sa) = 1 \;\; \forall s,a \tag{14}$$

This implicitly requires ϕ to be surjective, i.e. $\phi(\mathcal{H})=\mathcal{S}$, which can always be made true by defining $\mathcal{S} := \mathcal{S}_\phi := \phi(\mathcal{H})$. Note that the sum is taken over histories of any/mixed length. In general, B is a somewhat weird distribution, since it assigns probabilities to past and future observations given the current state and action. The interpretation and choice of B does not need to concern us, except later when we want to learn p.

The MDP requirement (6) will be replaced by the following definition:

$$\begin{aligned} p(s'r'|sa) &:= \sum_{h\in\mathcal{H}} P_\phi(s'r'|ha)B(h|sa) \\ &\equiv \sum_{t=1}^{\infty}\sum_{h_t\in\mathcal{H}_t} P_\phi(s_{t+1}=s', r_{t+1}=r'|h_t, a_t=a)B(h_t|sa) \end{aligned} \tag{15}$$

That is, the finite-state stationary MDP p is built from feature map ϕ, dispersion probability B, and environment P: The p-probability of observing state-reward pair (s',r') from state-action pair (s,a) is defined as the B-average over all histories h consistent with (s,a) of the P_ϕ-probability of observing (s',r') (obtained from P by ϕ-marginalizing) given history h and action a. The r.h.s. of the first line is merely shorthand for the second line. Note that $sas'r'$ are fixed and do not appear in h

which ranges over histories $\mathcal{H}$ of all lengths. It is easy to see that p is a probability distribution, and it is Markov by definition. If $P_\phi \in \text{MDP}$, then definition (15) coincides with p defined in (6). In general, the MDP p, depending on arbitrary B, is *not* the state distribution induced by P (and Π), which in general is non-Markov. Note that p is a stationary MDP for any B satisfying (14) and *any* ϕ and P. We need the following lemmas:

Some lemmas. The first lemma establishes the key relation between P and p via B used later to relate original history Bellman (optimality) equations (2–4) with reduced state Bellman (optimality) equations (7–9).

Lemma 3 (B-P-p relation) *For any function $f:\mathcal{S}\times\mathcal{R}\to\mathbb{R}$ and p defined in (15) in terms of P via (5), and $s':=\phi(h')$ and $h':=hao'r'$ it holds*

$$\sum_{h\in\mathcal{H}} B(h|sa)\sum_{o'r'} P(o'r'|ha)f(\underset{\substack{\uparrow\\ \text{depends on } hao'r'}}{s',r'}) \;=\; \sum_{s'r'} p(s'r'|sa)f(s',r')$$

Proof.

$$\begin{aligned}
&\sum_{h\in\mathcal{H}} B(h|sa)\sum_{o'r'} P(o'r'|ha)f(s',r')\\
\overset{(a)}{=}\;&\sum_{h\in\mathcal{H}} B(h|sa)\sum_{s'r'}\sum_{o':\phi(h')=s'} P(o'r'|ha)f(s',r')\\
\overset{(b)}{=}\;&\sum_{h\in\mathcal{H}} B(h|sa)\sum_{s'r'} P_\phi(s'r'|ha)f(s',r')\\
\overset{(c)}{=}\;&\sum_{s'r'} p(s'r'|sa)f(s',r')
\end{aligned}$$

In (a) we sum over all o' by first summing over all o' such that $\phi(hao'r')=s'$ and then summing over all s'. In (b) we used the definition (5) of P_ϕ. In (c) we used the definition (15) of p. ■

Inequalities (10) and (12) trivially bound $v-V$ differences in terms of $q-Q$ differences: $|v-V|\le\max_a|q-Q|$. The following lemma shows that a reverse holds in expectation, i.e. $|q-\langle Q\rangle_B|\le\gamma|v-V|$. The expectation can (only) be dropped if Q is constant for $h\in\phi^{-1}(s)$. Formally define

$$\langle f(h,a)\rangle_B \;:=\; \sum_{\tilde{h}\in\mathcal{H}} B(\tilde{h}|sa)f(\tilde{h},a), \quad \text{where} \quad s:=\phi(h) \tag{16}$$

That is, $\langle f(h,a)\rangle_B$ takes a B-average over all $\tilde{h}$ that ϕ maps to the same state as h. For convenience we will drop the tilde, which we can do if we declare $s:=\phi(h)$ to refer to the 'global' h in $\langle f(h,a)\rangle_B$ and not to the 'local' variable in the $h\in\mathcal{H}$ sum.

Lemma 4 ($|q-\langle Q\rangle|\leq\gamma|v-V|$) *For any P, ϕ, B, define p via* (15) *and* (5).
(i) If $|v^\pi(s)-V^\Pi(h)|\leq\delta\ \forall s=\phi(h)$
then $|q^\pi(s,a)-\langle Q^\Pi(h,a)\rangle_B|\leq\gamma\delta\ \forall s=\phi(h)\ \forall a$.
(ii) If $|v^*(s)-V^*(h)|\leq\delta\ \forall s=\phi(h)$
then $|q^*(s,a)-\langle Q^*(h,a)\rangle_B|\leq\gamma\delta\ \forall s=\phi(h)\ \forall a$.

Proof. (i) Let $s:=\phi(h)$ and $h':=hao'r'$ and $s':=\phi(h')$. Then

$$\begin{aligned}
\langle Q^\Pi(h,a)\rangle_B &\overset{(16)}{=} \sum_{h\in\mathcal{H}} B(h|sa)Q^\Pi(h,a)\\
&\overset{(2)}{=} \sum_{h\in\mathcal{H}} B(h|sa)\sum_{o'r'} P(o'r'|ha)[r'+\gamma V^\Pi(h')]\\
&\overset{(a)}{\leqslant} \sum_{h\in\mathcal{H}} B(h|sa)\sum_{o'r'} P(o'r'|ha)[r'+\gamma(v^\pi(s')\pm\delta)]\\
&\overset{Lem.3}{=} \sum_{s'r'} p(s'r'|sa)[r'+\gamma v^\pi(s')]\pm\gamma\delta\\
&\overset{(7)}{=} q^\pi(s,a)\pm\gamma\delta
\end{aligned}$$

In (a) we used the assumption (i) of the Lemma. The derived upper and lower bounds imply $|q^\pi(s,a)-\langle Q^\Pi(h,a)\rangle_B|\leq\gamma\delta$ (for all $s=\phi(h)$ and a).

(ii) follows the same steps except with Π and π replaced by Π^* and π^*, and using (3) and (8) instead of (2) and (7) to justify the steps. Note that in general $\Pi^*\neq\pi^*$! ■

5 Approximate Aggregation Results

This section contains the main technical contribution of the paper. We show that histories (or raw states) can be aggregated and modeled by an MDP even if the true aggregated process is actually not an MDP. A necessary condition for successful aggregation is of course that the quantities of interest, namely (Q-)Value functions and Policies can be represented as functions of the aggregated states. The results in this section roughly show that this necessary condition, which is significantly weaker than the MDP requirement, is also sufficient. All but one result also holds for approximate aggregation, i.e. approximate conditions lead to approximate reductions. We also lift the stationarity assumption.

- Theorem 5 shows how (approximately) ϕ-uniform Q^Π and Π can be obtained from the reduced Bellman equations (7).
- Theorem 6 weakens the assumptions and conclusions to (approximately) ϕ-uniform V^Π and Π.

- Theorem 7 shows that for (approximately) ϕ-uniform Q^*, the optimal policy is (approximately) ϕ-uniform, and (an approximation of it) can be obtained via the reduced Bellman optimality equations (8).
- Theorem 9 shows that for (approximately) ϕ-uniform V^* and Π^* we can obtain similar but somewhat weaker results. The proof of the latter involves extra complications not present in the other three proofs. Indeed, whether the arguably most desirable bound holds is Open Problem 10.

Note that all theorems crucially differ in their conditions and conclusions.

Theorem 5 ($\phi Q \pi$) *For any P, ϕ, and B, define p via (15) and (5). Let Π be some policy such that $\Pi(h)=\Pi(\tilde{h})$ and $|Q^\Pi(h,a)-Q^\Pi(\tilde{h},a)|\leq\varepsilon$ for all $\phi(h)=\phi(\tilde{h})$ and all a. Then for all a and h it holds:*

$$|Q^\Pi(h,a) - q^\pi(s,a)| \leq \frac{\varepsilon}{1-\gamma} \qquad \text{and} \qquad |V^\Pi(h) - v^\pi(s)| \leq \frac{\varepsilon}{1-\gamma},$$
$$\text{where} \quad \pi(s) := \Pi(h) \quad \text{and} \quad s = \phi(h)$$

Proof. Let $\delta := \sup_{s=\phi(h),a} |q^\pi(s,a) - Q^\Pi(h,a)|$. Then $|v^\pi(s) - V^\Pi(h)| \leq \delta \ \forall s = \phi(h)$ by (10),

$$\text{hence} \quad |q^\pi(s,a) - \langle Q^\Pi(h,a)\rangle_B| \leq \gamma\delta \quad \forall s = \phi(h), a$$

by Lemma 4i. By assumption on Q^Π and B, for $s=\phi(h)$ we have

$$\langle Q^\Pi(h,a)\rangle_B \equiv \sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s} B(\tilde{h}|sa) Q^\Pi(\tilde{h},a) \lesseqgtr \sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s} B(\tilde{h}|sa)[Q^\Pi(h,a) \pm \varepsilon] = Q^\Pi(h,a) \pm \varepsilon$$

Together this implies $|q^\pi(s,a) - Q^\Pi(h,a)| \leq \gamma\delta + \varepsilon$, hence $\delta \leq \gamma\delta + \varepsilon$, hence $\delta \leq \frac{\varepsilon}{1-\gamma}$. ■

Theorem 6 ($\phi V \pi$) *For any P, ϕ, and B, define p via (15) and (5). Let Π be some policy such that $\Pi(h)=\Pi(\tilde{h})$ and $|V^\Pi(h)-V^\Pi(\tilde{h})|\leq\varepsilon$ for all $\phi(h)=\phi(\tilde{h})$. Then for all a and h it holds:*

$$|V^\Pi(h) - v^\pi(s)| \leq \frac{\varepsilon}{1-\gamma} \qquad \text{and} \qquad |q^\pi(s,a) - \langle Q^\Pi(h,a)\rangle_B| \leq \frac{\varepsilon\gamma}{1-\gamma}$$
$$\text{where} \quad \pi(s) := \Pi(h) \quad \text{and} \quad s = \phi(h)$$

Proof. Let $\delta := \sup_{s=\phi(h),a} |v^\pi(s) - V^\Pi(h)|$, fix some $s=\phi(h)$, and let $a^\pi := \Pi(h)$. Now

$$\begin{aligned} \langle Q^\Pi(h,a^\pi)\rangle_B &\equiv \sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s} B(\tilde{h}|sa^\pi) Q^\Pi(\tilde{h},a^\pi) \overset{(a)}{=} \sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s} B(\tilde{h}|sa^\pi) V^\Pi(\tilde{h}) \\ &\lesseqgtr \sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s} B(\tilde{h}|sa^\pi)[V^\Pi(h) \pm \varepsilon] = V^\Pi(h) \pm \varepsilon \end{aligned} \tag{17}$$

where (a) follows from $a^\pi = \Pi(h) = \Pi(\tilde{h})$ and $Q^\Pi(\tilde{h},\Pi(\tilde{h})) = V^\Pi(\tilde{h})$. By Lemma 4i we have

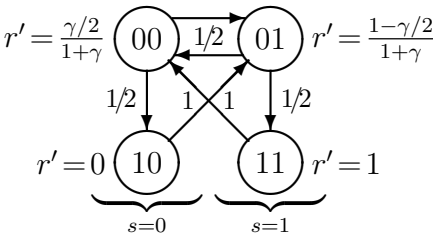
$$|q^\pi(s,a) - \langle Q^\Pi(h,a)\rangle_B| \leq \gamma\delta \quad \forall s = \phi(h), a \tag{18}$$

We also have $q^\pi(s,a^\pi) = q^\pi(s,\pi(s)) = v^\pi(s)$ from (7). Together with (17) and (18) for $a = a^\pi$ this yields

$$|v^\pi(s) - V^\Pi(h)| \;\leq\; |v^\pi(s) - \langle Q^\Pi(h,a^\pi)\rangle_B| + |\langle Q^\Pi(h,a^\pi)\rangle_B - V^\Pi(h)| \;\leq\; \gamma\delta + \varepsilon$$

hence $\delta \leq \gamma\delta + \varepsilon$ by the definition of δ, hence $\delta \leq \frac{\varepsilon}{1-\gamma}$. Note that while $|q^\pi(s,a^\pi) - Q^\Pi(h,a^\pi)| \leq \frac{\varepsilon\gamma}{1-\gamma}$, in general $|q^\pi(s,a) - Q^\Pi(h,a)| \not\leq \frac{\varepsilon\gamma}{1-\gamma}$ for $a \neq a^\pi$. ■

Example. Consider a process P which itself is an MDP in the observations with transition matrix T and reward function R, i.e. $P(o'r'|ha) = T^a_{oo'} R^{ar'}_{oo'}$. The example on the right has the special form $P(o'r'|ha) = T_{oo'} \cdot [\![r' = R(o)]\!]$. It is an action-independent Markov process T with deterministic reward function R, which can be read off from the diagram. Observation space is $\mathcal{O} = \{00,01,10,11\}$. Consider reduction



$$s_t \;:=\; \phi(h_t) \;:=\; \left\{\begin{array}{ll} 0 & \text{if } \; o_t = 00 \text{ or } 10 \\ 1 & \text{if } \; o_t = 01 \text{ or } 11 \end{array}\right\} \;\in\; \mathcal{S} \;:=\; \{0,1\}$$

The reduced process P_ϕ is not (even approximately) Markov:

$$\begin{array}{rcl} P_\phi(s'=0|o=00) & = & T_{00,00} + T_{00,10} \;=\; 0 \;+\; 1/2 \;=\; 1/2 \\ P_\phi(s'=0|o=10) & = & T_{10,00} + T_{10,10} \;=\; 0 \;+\; \;\;0 \;=\; \;\;0 \end{array} \quad \neq$$

That is, P violates the bisimulation condition [GDG03], and raw states 00 and 10 have a large bisimulation distance [FPP04, Ort07]. On the other hand, the (Q-)Value function $V(o_t) := V^\pi(h_t) = Q^\pi(h_t,a_t) \forall a_t$ can easily be verified to be

$$V(00) \;=\; V(10) \;=\; \frac{\gamma}{1-\gamma^2} \qquad \text{and} \qquad V(01) \;=\; V(11) \;=\; \frac{1}{1-\gamma^2}$$

That is, V and Q are ϕ-uniform. The conditions of Theorems 5 and 6 are satisfied exactly ($\varepsilon = 0$), and hence the four raw states $\mathcal{O}$ can be aggregated into two states $\mathcal{S}$ despite $P_\phi \notin$MDP (the policy is irrelevant and can be chosen constant). ♢

We now turn from the fixed policy case to similar theorems for optimal policies.

Theorem 7 ($\phi Q*$) *For any P, ϕ, and B, define p via (15) and (5). Assume $|Q^*(h,a)-Q^*(\tilde{h},a)| \leq \varepsilon$ for all $\phi(h)=\phi(\tilde{h})$ and all a. Then for all a and h and $s=\phi(h)$ it holds:*

$$(i) \quad |Q^*(h,a) - q^*(s,a)| \leq \frac{\varepsilon}{1-\gamma} \quad \textit{and} \quad |V^*(h) - v^*(s)| \leq \frac{\varepsilon}{1-\gamma},$$

$$(ii) \quad 0 \;\leq\; V^*(h) - V^{\tilde{\Pi}}(h) \;\leq\; \frac{2\varepsilon}{(1-\gamma)^2}, \quad \textit{where} \quad \tilde{\Pi}(h) := \pi^*(s)$$

$$(iii) \quad \textit{If } \varepsilon = 0 \textit{ then } \Pi^*(h) = \pi^*(s)$$

Proof. (i) The proof follows the same steps as the proof of Theorem 5, replacing all Π and π by $*$ and using (12) instead of (10) and Lemma 4ii instead of Lemma 4i to justify the steps.

(iii) If $\varepsilon=0$, then $Q^*(h,a)=q^*(s,a)$ by (i) implies $\Pi^*(h)=\pi^*(s)$, where it is worthwhile to carefully check that the latter has actually not been used inadvertently in proving the former. Cf. the next theorem and proof.

(ii) For $s=\phi(h)$ and $\tilde{a}:=\tilde{\Pi}(h)=\pi^*(s)$,

$$V^*(h) - \frac{\varepsilon}{1-\gamma} \overset{(i)}{\leq} v^*(s) \overset{(8)}{=} q^*(s,\tilde{a}) \overset{(i)}{\leq} Q^*(h,\tilde{a}) + \frac{\varepsilon}{1-\gamma}$$

which implies $Q^*(h,\tilde{\Pi}(h)) \geq V^*(h) - \frac{2\varepsilon}{1-\gamma}$. The claim now follows from the next Lemma 8 below. ■

The following lemma shows that if replacing the first action after h of the optimal policy Π^* by the action provided by Π thereafter following Π^* is at most ε-suboptimal, then always using Π is at most $\frac{\varepsilon}{1-\gamma}$-suboptimal.

Lemma 8 ($Q\pi*$) *If $Q^*(h,\Pi(h)) \geq V^*(h) - \varepsilon$ for all h for some policy Π, then for all h and a*

$$0 \;\leq\; Q^*(h,a) - Q^{\Pi}(h,a) \;\leq\; \frac{\varepsilon\gamma}{1-\gamma} \qquad \textit{and} \qquad 0 \;\leq\; V^*(h) - V^{\Pi}(h) \;\leq\; \frac{\varepsilon}{1-\gamma}$$

Proof. Let $\delta := \sup_{h,a}[Q^*(h,a) - Q^{\Pi}(h,a)]$. This implies

$$0 \overset{(a)}{\leq} V^*(h) - V^{\Pi}(h) \overset{(b)}{\leq} \varepsilon + Q^*(h,\Pi(h)) - Q^{\Pi}(h,\Pi(h)) \overset{(c)}{\leq} \varepsilon + \delta \qquad (19)$$

(a) follows from (1); (b) by assumption; and (c) by definition of δ for $a=\Pi(h)$. Now for any a and h, this implies

$$\begin{aligned} Q^{\Pi}(h,a) &\overset{(1)}{\leq} Q^*(h,a) \overset{(3)}{=} \sum_{o'r'} P(o'r'|ha)[r' + \gamma V^*(h')] \qquad [h' = hao'r'] \\ &\overset{(19)}{\leq} \sum_{o'r'} P(o'r'|ha)[r' + \gamma(V^{\Pi}(h') + \varepsilon + \delta)] \overset{(2)}{=} Q^{\Pi}(h,a) + \gamma(\varepsilon + \delta) \end{aligned}$$

Hence $\delta \leq \gamma(\varepsilon+\delta)$, hence $\delta \leq \frac{\varepsilon\gamma}{1-\gamma}$. ■

Theorem 9 ($\phi V*$) *For any P, ϕ, and B, define p via (15) and (5). Assume $\Pi^*(h)=\Pi^*(\tilde{h})$ and $|V^*(h)-V^*(\tilde{h})|\leq\varepsilon$ for all $\phi(h)=\phi(\tilde{h})$. Then for all a and h and $s=\phi(h)$ it holds:*

$$(i)\quad |V^*(h)-v^*(s)|\leq\frac{3\varepsilon}{(1-\gamma)^2}\quad \textit{and}\quad |q^*(s,a)-\langle Q^*(h,a)\rangle_B|\leq\frac{3\varepsilon\gamma}{(1-\gamma)^2},$$
$$(ii)\quad \textit{If } \varepsilon=0 \textit{ then } \Pi^*(h)=\pi^*(s)$$

The proof actually implies the stronger lower bound $V^*(h)-v^*(s)\geq\frac{3\varepsilon}{1-\gamma}$ and similarly for Q^*, but we do not know whether the upper bound can be improved.

Proof. While proofs start to get routine, here is a warning that care is in order when recycling similar proofs. Theorem 6 relies on the assumption that $\pi(s)=\Pi(h)$ for $s=\phi(h)$, while we were lucky that the proof of Theorem 7 worked without knowing $\pi^*(s)=\Pi^*(h)$ in advance. Here we have to work a bit harder.

Let us define $a^0:=\pi^0(s):=\Pi^*(h)$ for $s=\phi(h)$. The Bellman equation for policy π^0 is

$$q^{\pi^0}(s,a)\;=\;\sum_{s'r'}p(s'r'|sa)[r'+\gamma v^{\pi^0}(s')]\quad\text{and}\quad v^{\pi^0}(s)=q^{\pi^0}(s,\pi^0(s)) \tag{20}$$

At this stage π^0 may well be different from π^*, since π^* satisfies (8), not (20), but we will now show that it actually does. First note that

$$q^{\pi^0}(s,a^0)\;=\;v^{\pi^0}(s)\;\leqslant\;V^{\Pi^*}(h)\pm\frac{\varepsilon}{1-\gamma}\;=\;V^*(h)\pm\frac{\varepsilon}{1-\gamma} \tag{21}$$

where the bounds follow from Theorem 6 applied to $\Pi:=\Pi^*$ (with $\pi=\pi^0$). For general a we only get an upper bound:

$$\begin{aligned}q^{\pi^0}(s,a)-\frac{\varepsilon\gamma}{1-\gamma}\;&\overset{Thm.6}{\leq}\;\langle Q^{\Pi^*}(h,a)\rangle_B\;\overset{(16)}{=}\;\sum_{h\in\mathcal{H}}B(h|sa)Q^*(h,a)\\&\overset{(4)}{\leq}\;\sum_{h\in\mathcal{H}}B(h|sa)Q^*(h,\Pi^*(h))\;\underset{(3)}{\overset{(14)}{=}}\;\sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s}B(\tilde{h}|sa)V^*(\tilde{h})\\&\overset{(a)}{\leq}\;\sum_{\tilde{h}\in\mathcal{H}:\phi(\tilde{h})=s}B(\tilde{h}|sa)[V^*(h)+\varepsilon]\;\overset{(14)}{=}\;V^*(h)+\varepsilon\end{aligned} \tag{22}$$

(a) uses the theorem's assumption on $V^*(h)$. Together, (21) and (22) imply

$$v^{\pi^0}(s)\;\overset{(20)}{=}\;q^{\pi^0}(s,a^0)\;\leq\;\max_a q^{\pi^0}(s,a)\;\overset{(22)}{\leq}\;V^*(h)+\frac{\varepsilon}{1-\gamma}\;\overset{(21)}{\leq}\;v^{\pi^0}(s)+\frac{2\varepsilon}{1-\gamma} \tag{23}$$

(ii) For $\varepsilon=0$, the previous equation implies $v^{\pi^0}(s)=\max_a q^{\pi^0}(s,a)$, hence (20) can be rewritten as

$$q^{\pi^0}(s,a)\;=\;\sum_{s'r'}p(s'r'|sa)[r'+\gamma v^{\pi^0}(s')]\quad\text{and}\quad v^{\pi^0}(s)=\max_a q^{\pi^0}(s,a)$$

This shows that (q^{π^0},v^{π^0}) satisfies the same Bellman *optimality* equation as (q^*,v^*) does. Since it has a unique solution, we must have $q^{\pi^0}\equiv q^*$ and $v^{\pi^0}\equiv v^*$ and $\pi^*\equiv\pi^0$, which for $s=\phi(h)$ implies $\Pi^*(h)=\pi^*(s)$ by definition of π^0. It also implies $V^*(h)=v^*(s)$ by (21), and $q^*(s,a)=\langle Q^*(h,a)\rangle_B$ by Lemma 4ii, i.e. the $\varepsilon=0$ version of (i).

(i) We now continue with the general $\varepsilon>0$ case. For all s and a we have

$$0 \overset{(9)}{\leq} q^*(s,a)-q^{\pi^0}(s,a) \underset{(8)}{\overset{(7)}{=}} \sum_{s'r'} p(s'r'|sa)\gamma(v^*(s')-v^{\pi^0}(s')) \overset{(a)}{\leq} \gamma\max_{s'}\{v^*(s')-v^{\pi^0}(s')\}$$

$$0 \overset{(9)}{\leq} v^*(s)-v^{\pi^0}(s) \underset{(8)}{\overset{(23)}{\leq}} \max_a q^*(s,a)-\max_a q^{\pi^0}(s,a)+\tfrac{2\varepsilon}{1-\gamma} \overset{(13)}{\leq} \max_a\{q^*(s,a)-q^{\pi^0}(s,a)\}+\tfrac{2\varepsilon}{1-\gamma}$$

In (a) we have upper bounded the p-expectation by the maximum. Together this gives

$$\begin{aligned} \max_s\{v^*(s)-v^{\pi^0}(s)\} &\leq \gamma\max_s\{v^*(s)-v^{\pi^0}(s)\}+\frac{2\varepsilon}{1-\gamma} \\ \Rightarrow \quad \max_s\{v^*(s)-v^{\pi^0}(s)\} &\leq \frac{2\varepsilon}{(1-\gamma)^2} \end{aligned}$$

$$\text{Hence for } s=\phi(h):\ V^*(h)-\frac{\varepsilon}{1-\gamma} \overset{(21)}{\leq} v^{\pi^0}(s) \overset{(9)}{\leq} v^*(s) \overset{\nwarrow}{\leq} v^{\pi^0}(s)+\frac{2\varepsilon}{(1-\gamma)^2} \tag{24}$$

$$\overset{(21)}{\leq} V^*(h)+\frac{\varepsilon}{1-\gamma}+\frac{2\varepsilon}{(1-\gamma)^2} \leq V^*(h)+\frac{3\varepsilon}{(1-\gamma)^2}$$

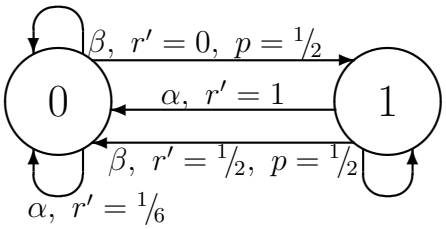
Together with Lemma 4ii this implies (i). ■

We are primarily interested in the optimal policy $\Pi^*(h)$; to correctly represent the value $V^*(h)$ is only of indirect interest. If Π^* is ϕ-uniform, it can be represented as $\Pi^*(h)=\pi^0(s)$ for some π^0, but if the ϕ-uniformity condition on V^* in Theorem 9 is dropped, the conclusion $\Pi^*(h)=\pi^*(s)$ can fail as the following example shows.

Counter Example. Let P be the MDP $P(o'r'|ha):=T^a_{oo'}\cdot[\![r'=R^a_o]\!]$ with two raw states $o\in\{0,1\}$ and two actions $a\in\{\alpha,\beta\}$ formally defined on the left and depicted on the right:

$$T^\alpha := \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad R^\alpha := \begin{pmatrix} 1/6 \\ 1 \end{pmatrix},$$

$$T^\beta := \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \qquad R^\beta := \begin{pmatrix} 0 \\ 1/2 \end{pmatrix},$$



The value of policy π in vector notation is $V^\pi=R^\pi+\gamma T^\pi V^\pi$, where $V^\pi_{o_t}:=V^\pi(h_t)$. The 4 stationary policies are denoted by $\pi=a^0a^1$, where a^o is the action taken in raw state o. For $\gamma=0$, their values are

$\gamma=0,\ \pi$	$\alpha\alpha$	$\alpha\beta$	$\beta\alpha$	$\beta\beta$
$V^\pi_0=R^{\pi(0)}_0$	1/6	1/6	0	0
$V^\pi_1=R^{\pi(1)}_1$	1	1/2	1	1/2

Policy $\pi=\alpha\alpha$ has the highest value, therefore $\Pi^*(h)\equiv\alpha$. Let us now aggregate raw states $o\in\{0,1\}$ to a 1-state MDP. Its value is $v=\frac{1}{1-\gamma}\rho^\top R=\rho^\top R$ for $\gamma=0$, where ρ is the stationary distribution $\rho^\top=\rho^\top T$ of T, in particular $\rho^\alpha=\binom{1}{0}$ for T^α and $\rho^\beta=\binom{1/2}{1/2}$ for T^β. Since there is only 1 aggregated state, there are only 2 stationary policies, one for each action. This leads to $v^\alpha=\frac{1}{6}<\frac{1}{4}=v^\beta$, hence $\pi^*(s)\equiv\beta\neq\alpha\equiv\Pi^*(h)\ \forall s,h$. That is, despite Π^* being constant, $\pi^*\neq\Pi^*$, which shows that the condition on V^* in Theorem 9 cannot be dropped. Note that $V^*=V^{\alpha\alpha}=\binom{1/6}{1}$ is far from constant. By continuity, the policy reversal also holds for $\gamma>0$. Indeed, this example works for all $\gamma<\frac{2}{5}$ and other examples work for all $0\leq\gamma<1$. ♢

Open Problem 10 ($\phi V*$) *Under the same conditions as Theorem 9, is*

$$V^*(h)-V^{\tilde\Pi}(h)\ \overset{??}{=}\ O\Big(\frac{\varepsilon}{(1-\gamma)^{?}}\Big)\quad \textit{where}\quad \tilde\Pi(h):=\pi^*(s) \tag{25}$$

Arguments. Here are some arguments why it might be true (or false):

(1) For $\varepsilon=0$ it immediately follows from Theorem 9, since in this case $\tilde\Pi(h)=\Pi^*(h)$. Some continuity argument might allow to establish a bound for small $\varepsilon>0$.

(2) Theorem 7i&iii mostly carried over to Theorem 9, so a-priori it is not too implausible that Theorem 7ii carries over to (25). On the other hand, the proofs of (i) and (iii) of both theorems were sufficiently different, so the analogy argument is weak.

(3) Let $\tilde a:=\tilde\Pi(h):=\pi^*(s)$ for $s=\phi(h)$. Then

$$\begin{array}{lllllll}\langle Q^*(h,\tilde a)\rangle_B & \overset{Thm.9i}{\geq} & q^*(s,\tilde a)-\frac{3\varepsilon\gamma}{(1-\gamma)^2} & \overset{(8)}{=} & v^*(s)-\frac{3\varepsilon\gamma}{(1-\gamma)^2} & \overset{(24)}{\geq} & V^*(h)-\frac{3\varepsilon}{(1-\gamma)^2}\\ Q^*(h,\tilde a) & \overset{(3)}{\leq} & V^*(h) & & & & \end{array}$$

For $\varepsilon=0$ this pair of inequalities implies that $Q^*(h,\tilde a)$ lower bounds its own expectation, therefore it must be constant and equal to $V^*(h)$ on each $\phi^{-1}(s)$-partition. For $\varepsilon>0$, with high probability $Q^*(h,\tilde a)$ cannot be much smaller than $V^*(h)$. If it weren't for the probability qualifier we could now apply Lemma 8 to establish (25) (as in the proof of Theorem 7ii). Low probability events could invalidate this argument. ■

Discussion. Open Problem 10 would be the main result if we had a proof for $\varepsilon>0$. Absent of it we have to be content with Theorem 7ii. Both statements imply that we can aggregate histories as much as we wish, as long as the optimal value function and policy are still approximately representable as functions of aggregated states. Whether the reduced process P_ϕ is Markov or not is immaterial. We can use surrogate MDP p to find an ε-optimal policy for P.

Most RL work, including on state aggregation, is formulated in terms of MDPs, i.e. the original process P is already an MDP. Let us call this the original or raw MDP. We could interpret the whole history as a raw state, which formally makes every P an MDP, but normally only observations are identified with raw states, i.e. P is a raw MDP iff $P(o'r'|ha)=P(o'r'|oa)$. In this case, $V^*(h_t)=V^*(o_t)$ etc. depends on raw states only (which is well known or follows from Theorem 2 with $\phi(h_t)=o_t$). Since our results hold for all P, they clearly hold if P is a raw MDP and if $\phi(h_t):=\phi(o_t)$ maps raw states to aggregated states.

The remainder of this paper shows how much we can aggregate and how to develop RL algorithms exploiting these insights.

6 Extreme Aggregation

The results of Section 5 showed that histories can be aggregated and modeled by an MDP even if the true aggregated process is not an MDP. The only restrictions were that the (Q-)Value functions and Policies could still be (approximately) represented as functions of the aggregated states. We will see in this section that in theory this allows to represent *any* process P as a small finite-state MDP.

Extreme aggregation based on Theorem 7. Consider ϕ that maps each history to the vector-over-actions of optimal Q-values $Q^*(h,\cdot)$ discretized to some finite ε-grid:

$$\phi(h) \;:=\; \big(\lfloor Q^*(h,a)/\varepsilon \rfloor\big)_{a\in\mathcal{A}} \;\in\; \{0,1,...,\lfloor\tfrac{1}{\varepsilon(1-\gamma)}\rfloor\}^{\mathcal{A}} \;=:\; \mathcal{S} \tag{26}$$

That is, all histories with ε-close Q^*-values are mapped to the same state:

$$|Q^*(h,a)-Q^*(\tilde{h},a)| \;\leq\; \varepsilon \quad \forall\phi(h)=\phi(\tilde{h})\;\forall a$$

Now choose some B and determine p from P via (15) and (5). Find the optimal policy π^* of MDP p of size $|\mathcal{S}|$. Define $\tilde{\Pi}(h):=\pi^*(\phi(h))$. By Theorem 7ii, $\tilde{\Pi}$ is an ε'-optimal policy of original process P in the sense that

$$|V^{\tilde{\Pi}}(h)-V^*(h)| \;\leq\; \frac{2\varepsilon}{(1-\gamma)^2} \;=:\; \varepsilon'$$

Extreme aggregation based on Open Problem 10. If (25) holds, we can aggregate even better: Consider ϕ that maps each history to the optimal Value $V^*(h)$ discretized to some finite ε-grid and to the optimal action $\Pi^*(h)$:

$$\phi(h) \;:=\; \big(\lfloor V^*(h)/\varepsilon \rfloor, \Pi^*(h)\big) \;\in\; \{0,1,...,\lfloor\tfrac{1}{\varepsilon(1-\gamma)}\rfloor\}\times\mathcal{A} \;=:\; \mathcal{S} \tag{27}$$

That is, all histories with ε-close V^*-Values and same optimal action are mapped to the same state:

$$|V^*(h)-V^*(\tilde{h})| \;\leq\; \varepsilon \quad\text{and}\quad \Pi^*(h)=\Pi^*(\tilde{h}) \qquad \forall\phi(h)=\phi(\tilde{h})$$

As before, determine p, find its optimal policy π^*, and define $\tilde{\Pi}(h) := \pi^*(\phi(h))$. If (25) holds, then $\tilde{\Pi}$ is an ε'-optimal policy of original process P in the sense that

$$|V^{\tilde{\Pi}(h)} - V^*(h)| \;=\; O\Big(\frac{\varepsilon}{(1-\gamma)^?}\Big) \;=:\; \varepsilon'$$

The following theorem summarizes the considerations for the two choices of ϕ above:

Theorem 11 (Extreme ϕ) *For every process P there exists a reduction ϕ ((26) or (27) will do) and MDP p defined via (15) and (5) whose optimal policy π^* is an ε'-optimal policy $\tilde{\Pi}(h) := \pi^*(\phi(h))$ for P. The size of the MDP is bounded (uniformly for* any *P) by*

$$|\mathcal{S}| \leq \Big(\frac{3}{\varepsilon'(1-\gamma)^3}\Big)^{|\mathcal{A}|} \quad \textit{and if (25) holds even by} \quad |\mathcal{S}| = O\Big(\frac{|\mathcal{A}|}{\varepsilon'(1-\gamma)^{1+?}}\Big)$$

Proof. For $\mathcal{S}$ defined in (26) we have

$$|\mathcal{S}| \;=\; \big(\lfloor\tfrac{1}{\varepsilon(1-\gamma)}\rfloor+1\big)^{|\mathcal{A}|} \;=\; \big(\lfloor\tfrac{2}{\varepsilon'(1-\gamma)^3}\rfloor+1\big)^{|\mathcal{A}|} \;\leq\; \big(\tfrac{3}{\varepsilon'(1-\gamma)^3}\big)^{|\mathcal{A}|}$$

where in the last inequality we have assumed $\varepsilon' \leq \frac{1}{1-\gamma}$. (For $\varepsilon' > \frac{1}{1-\gamma}$ the theorem is trivial, since any policy is ε'-optimal). For $\mathcal{S}$ defined in (27) the derivation is similar. The theorem now follows from the considerations in the paragraphs before the theorem. ■

Discussion. A valid question is of course whether Theorem 11 is just an interesting theoretical insight/curiosity or of any practical use. After all, ϕ depends on Q^* (or V^* and Π^*), but if we knew Q^*, Π^* would readily be available and the detour through p and π^* pointless.

Theorem 11 reaches relevance by the following observation: If we start with a sufficiently rich class of maps Φ that contains at least one ϕ approximately representing $Q^*(h,\cdot)$, and have a learning algorithm that favors such ϕ, then Theorems 5–9 tell us that we do not need to worry about whether P_ϕ is MDP or not; we "simply" use/learn MDP p instead. Theorem 11 tells us that this allows for extreme aggregation far beyond MDPs.

This program is in parts worked out in the next two sections, but more research is needed for its completion. Learning p from (real) P-samples is considered in Section 7 and learning ϕ in Section 8.

7 Reinforcement Learning

In RL, P and therefore p are unknown. We now show how to learn p from samples from P. For this we have to link B to the distribution over histories induced by P and to the behavior policy Π_B the agent follows. We still assume ϕ is given.

Behavior policy Π_B. Let $\Pi_B:\mathcal{H}\rightsquigarrow\mathcal{A}$ be the behavior policy of our RL agent, which in general is non-stationary due to learning, often stochastic to ensure exploration, and (usually) different from any policy considered so far ($\Pi^*,\pi^*,\tilde{\Pi},\pi^0,\Pi,\pi$). Note that a sequence of policies $\Pi_1,\Pi_2,...$ where each Π_t is learnt from h_t and used at time t (or for some number of steps) is nothing but a single non-stationary policy $\Pi_B(h_t)=\Pi_t(h_t)\forall t,h_t$, so Π_B indeed includes the case of policy learning.

Choice of B. The interaction of agent Π_B with environment P stochastically generates some history h_t followed by action a_t with joint probability, say $P_B(h_ta_t)$. We use subscripts B and/or ϕ to indicate dependence on Π_B and/or ϕ. A natural choice for $B(h|sa)$ in (14) would be to condition of P_B on s_ta_t. We now show that this does not work and how to fix the problem. We can get $P_{\phi B}(h_t|s_ta_t)$ from P and Π_B and several other useful distributions as follows:

$$\begin{aligned}
P_B(h_{t+1}|h_t) &= P(o_{t+1}r_{t+1}|h_ta_t)\Pi_B(a_t|h_t) \qquad [h_{t+1}=h_ta_to_{t+1}r_{t+1}]\\
P_B(h_n) &= \prod_{t=0}^{n-1} P_B(h_{t+1}|h_t), \qquad P_B(h_ta_t) = \Pi_B(a_t|h_t)P_B(h_t)\\
P_{\phi B}(s_ta_t) &= \sum_{h_t:\phi(h_t)=s_t} P_B(h_ta_t), \qquad P_{\phi B}(h_t|s_ta_t) = \frac{P_B(h_ta_t)}{P_{\phi B}(s_ta_t)}[\![\phi(h_t)=s_t]\!]\\
P_{\phi B}(s_{t+1}r_{t+1}|s_ta_t) &= \sum_{h_t:\phi(h_t)=s_t} P_\phi(s_{t+1}r_{t+1}|h_ta_t)P_{\phi B}(h_t|s_ta_t) \quad \text{[see (5) for def. of } P_\phi\text{]} \quad (28)\\
P_{\phi B}(s_ta_ts_{t+1}r_{t+1}) &= P_{\phi B}(s_{t+1}r_{t+1}|s_ta_t)P_{\phi B}(s_ta_t)
\end{aligned}$$

$P_{\phi B}(h_t|s_ta_t)$ has the following properties:

$$P_{\phi B}(h_t|s_ta_t)\geq 0 \quad \text{and} \quad \sum_{h_t\in\mathcal{H}_t} P_{\phi B}(h_t|s_ta_t) = \sum_{h_t:\phi(h_t)=s_t} P_{\phi B}(h_t|s_ta_t) = 1 \quad \forall t,s_t,a_t \quad (29)$$

This is close to the required condition (14) for B but crucially different. The sum in (14) is over histories of all lengths while in (29) the sum is limited to histories of length t. It is easy to miss this difference due to the compact notation. Technically P_B is a probability measure on infinite sequences $\mathcal{H}_\infty$ and $P_B(h_t)$ is short for $P_B(\Gamma_{h_t})$ where Γ_{h_t} is the set of infinite histories starting with h_t, i.e. $P_B(h_t)$ is the probability that the infinite history starts with h_t ($\sum_{h_t\in\mathcal{H}_t}P_B(h_t)=1\forall t$). On the other hand, $B(h)$ is a probability distribution over finite histories of mixed length ($\sum_{h\in\mathcal{H}}B(h)=1$); similarly for P_B and B conditioned on / parameterized by s and a.

We can fix this mismatch by introducing weights $w_t:\mathcal{S}\times\mathcal{A}\rightsquigarrow[0;1]$ and define

$$B(h_t|sa) := w_t(sa)P_{\phi B}(h_t|s_t=s,a_t=a)\ \forall t, \quad \text{where} \quad \sum_{t=1}^{\infty} w_t(sa)=1\ \forall s,a \quad (30)$$

which now satisfies (14) (due to $\sum_{h\in\mathcal{H}}=\sum_{t=1}^{\infty}\sum_{h_t\in\mathcal{H}_t}$). MDP p can now be repre-

sented as

$$\begin{aligned} p(s'r'|sa) &= \sum_{t=1}^{\infty} w_t(sa) \sum_{h_t \in \mathcal{H}_t} P_\phi(s_{t+1}=s', r_{t+1}=r'|h_t, a_t=a) P_{\phi B}(h_t|s_t=s, a_t=a) \\ &= \sum_{t=1}^{\infty} w_t(sa) P^t_{\phi B}(s'r'|sa) \end{aligned} \tag{31}$$

That is, p is the w-weighted time-average of $P^t_{\phi B}$. The first equality follows from (15) and (30); the second one from (28). We also introduced the shorthand $P^t_{\phi B}(s'r'|sa) := P_{\phi B}(s_{t+1}=s', r_{t+1}=r'|s_t=s, a_t=a)$.

Choice of w_t. If $P^t_{\phi B}$ in (31) is stationary, i.e. independent of t, then $p(s'r'|sa) = P^t_{\phi B}(s'r'|sa)$ for all t, since the weights sum to one, and estimation is easy. Note that in general we cannot estimate non-stationary $P^t_{\phi B}$, since for each t we have only one sample available, but we will see that estimation of p is still possible. Assume we have observed h_n, and choose

$$w_t(sa) \;:=\; \frac{P^t_{\phi B}(sa)}{\sum_{t=1}^n P^t_{\phi B}(sa)} \quad \text{for } t \le n \quad \text{and} \quad 0 \quad \text{for } t > n \tag{32}$$

Inserting this into (31) and using (28) gives

$$p(s'r'|sa) \;=\; \frac{\frac{1}{n}\sum_{t=1}^n P^t_{\phi B}(sas'r')}{\frac{1}{n}\sum_{t=1}^n P^t_{\phi B}(sa)} \tag{33}$$

We estimate numerator and denominator separately.

Law of large numbers. For $t=1,2,3,...$ let $X_t \in \{0,1\}$ be binary random variables with expectation $\mathbb{E}[X_t]$. Define $n_1 = \sum_{t=1}^n X_t$ be the number of sampled 1s. The strong law of large numbers says that

$$\frac{n_1}{n} - \frac{1}{n}\sum_{t=1}^n \mathbb{E}[X_t] \;\stackrel{n\to\infty}{\longrightarrow}\; 0 \quad \text{almost surely} \qquad \text{under weak conditions} \tag{34}$$

Note that the law holds far beyond i.i.d. random variables under a variety of conditions [Faz06, VGS05] which we collectively call 'weak conditions'. It is not even necessary for n_1/n to converge.

Estimation of p. Now fix some (s,a), and let $X_t := [\![s_t=s, a_t=a]\!]$. (Here we assume that variables in h_t are random variables and $sas'r'$ are realizations.) Then

$$n(sa) \;:=\; n_1 \;=\; \sum_{t=1}^n X_t \;=\; \#\{t \le n : s_t = s, a_t = a\}$$

is the number of times action a is taken in state s, and $\mathbb{E}[X_t] = P(X_t=1) = P^t_{\phi B}(sa)$, hence (34) implies

$$\frac{n(sa)}{n} - \frac{1}{n}\sum_{t=1}^n P^t_{\phi B}(sa) \;\stackrel{n\to\infty}{\longrightarrow}\; 0 \quad \text{a.s. under weak conditions} \tag{35}$$

Similarly for $Y_t := [\![s_t a_t s_{t+1} r_{t+1} = sas'r']\!]$ and $n(sas'r') := \sum_{t=1}^n Y_t$ we have

$$\frac{n(sas'r')}{n} - \frac{1}{n}\sum_{t=1}^{n} P^t_{\phi B}(sas'r') \;\stackrel{n\to\infty}{\longrightarrow}\; 0 \quad \text{with } P\text{-probability 1} \tag{36}$$

under weak conditions. (35) and (36) via (33) are nearly sufficient to imply

$$\frac{n(sas'r')}{n(sa)} - p(s'r'|sa) \;\stackrel{n\to\infty}{\longrightarrow}\; 0 \quad \text{almost surely} \tag{37}$$

A sufficient but by far not necessary condition is

$$\liminf_{n\to\infty} \frac{n(sa)}{n} \;>\; 0 \quad \text{almost surely} \tag{38}$$

Theorem 12 (p-estimation) *For B defined in* (30) *and* (32) *we have: If* (36) *and* (38) *hold, then* (37) *holds. For example, if Y_t are stationary ergodic processes, then* (36) *and* (38) *hence* (37) *hold for all state-action pairs that matter (i.e. for those occurring with non-zero probability).*

Proof. We introduce the following (n-dependent) shorthands:

$$\bar{X} := \frac{n(sa)}{n}, \qquad \bar{x} := \frac{1}{n}\sum_{t=1}^{n} P^t_{\phi B}(sa), \qquad \alpha := \liminf_{n\to\infty}\frac{n(sa)}{n},$$
$$\bar{Y} := \frac{n(sas'r')}{n}, \qquad \bar{y} := \frac{1}{n}\sum_{t=1}^{n} P^t_{\phi B}(sas'r')$$

With these abbreviations, assumption (36) implies (35), i.e.

$$\bar{Y} - \bar{y} \;\to\; 0 \quad \text{implies} \quad \bar{X} - \bar{x} \;=\; \sum_{s'r'} \bar{Y} - \sum_{s'r'} \bar{y} \;=\; \sum_{s'r'} [\bar{Y} - \bar{y}] \;\to\; 0 \tag{39}$$

since $\mathcal{S}$ and $\mathcal{R}$ have been assumed finite. Now

$$\Big|\frac{n(sas'r')}{n(sa)} - p(s'r'|sa)\Big| \;=\; \Big|\frac{\bar{Y}}{\bar{X}} - \frac{\bar{y}}{\bar{x}}\Big| \;\le\; \Big|\frac{\bar{Y}}{\bar{X}} - \frac{\bar{y}}{\bar{X}}\Big| \;+\; \Big|\frac{\bar{y}}{\bar{X}} - \frac{\bar{y}}{\bar{x}}\Big|$$
$$= \;\frac{1}{\bar{X}}|\bar{Y} - \bar{y}| \;+\; \frac{\bar{y}}{\bar{X}\bar{x}}|\bar{x} - \bar{X}| \;\le\; \frac{1}{\bar{X}}\Big(|\bar{Y} - \bar{y}| \;+\; |\bar{x} - \bar{X}|\Big) \;\stackrel{n\to\infty}{\longrightarrow}\; 0 \quad \text{a.s.}$$

The first inequality is just the triangle inequality. The second inequality follows from $\bar{y} \le \bar{x}$. The limit is zero, since almost surely $\limsup_{n\to\infty}[1/\bar{X}] = 1/\alpha < \infty$ and $\bar{Y} - \bar{y} \to 0$ and $\bar{X} - \bar{x} \to 0$. Hence (37) holds. Finally, for stationary ergodic Y_t, we have $\bar{y} = \frac{1}{n}\sum_{t=1}^n \mathbb{E}[Y_t] = \mathbb{E}[Y_1] = \textit{constant}$, and hence $\bar{x} = \sum_{s'r'} \bar{y} = \textit{constant}$. Therefore

$$\text{(36) holds by} \quad \bar{Y} \;=\; \frac{1}{n}\sum_{t=1}^{n} Y_t \;\stackrel{\text{ergodicity}}{\longrightarrow}\; \mathbb{E}[Y_1] \;\stackrel{\text{stationarity}}{=}\; \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[Y_t] \;=\; \bar{y},$$
$$\text{(38) holds by} \quad \liminf_{n\to\infty} \bar{X} \;\stackrel{(39)}{=}\; \bar{x} \;\stackrel{\text{stationarity}}{=}\; \mathbb{E}[X_1] \;=\; P^1_{\phi B}(sa) \;\stackrel{\text{assumption}}{>}\; 0$$

■

Discussion. Limit (37) shows that standard frequency estimation for p will converge to the true p under weak conditions. If P_ϕ is MDP, samples are conditionally i.i.d. and the 'weak conditions' are satisfied. But the law of large numbers and hence (37) holds far beyond the i.i.d. case [FK01, VGS05], e.g. for stationary ergodic processes. Condition (38) that every state-action pair be visited with non-vanishing relative frequency can be significantly relaxed. Stationarity is also not necessary, and indeed often does not hold due to a non-stationary environment P or a non-stationary behavior policy Π_B (or both).

Other choices for w_t are possible, e.g. we could multiply numerator and denominator of (32) by some arbitrary positive function $u_t(as)$, which leads to a weighted average estimator.

We estimate p in order to estimate q^* and ultimately π^*. This is model-based RL. We can also learn π^* model-free. For instance, condition (37) should be sufficient for Q-learning to converge to Q^*.

Q-learning and other RL algorithms designed for MDPs have been observed to often (but not always) perform well even if applied to non-MDP domains. Our results appear to explain why, but this calls for further investigations.

8 Feature Reinforcement Learning

The idea of FRL is to *learn* ϕ [Hut09c]. FRL starts with a class of maps Φ, compares different $\phi \in \Phi$, and *selects* the most appropriate one given the experience h_t so far. Several criteria based on how well ϕ reduces P to an MDP have been devised [Hut09b, Hut09a] and theoretically [SH10] and experimentally [NSH11] investigated [Ngu13]. Theorems 5–9 show that demanding P_ϕ to be approximately MDP is overly restrictive. Theorem 11 suggests that if we relax this condition, much more substantial aggregation is possible, provided Φ is rich enough.

(F)RL deals with the case of unknown P. We first discuss learning ϕ for the unrealistic case of exact aggregation ($\varepsilon=0$) and infinite sample size ($n=\infty$). This serves as a useful guide to work out its generalization to the realistic but significantly more complex case of approximate aggregation based on finite sample size. Finally we discuss a family of recent algorithms (BLB and extensions [Ngu13]) that appear to nearly have the right properties for our purpose. This section is more a collection of ideas and outlook towards algorithms exploiting and motivating the usefulness of the new insights obtained in the previous sections.

Search for exact ϕ based on infinite sample size. Since we are now concerned with comparing different $\phi \in \Phi$, we subscribe quantities with ϕ when necessary. Consider the unrealistic case of infinite sample size ($n=\infty$) and a search for exact reductions ϕ. We call a reduction $\phi:\mathcal{H}\to\mathcal{S}_\phi$ exact iff $Q^*(h,a)=q^*_\phi(s,a)$ and $\Pi^*(h)=\pi^*_\phi(s)$ for all $s=\phi(h)$ and a.

Even for $n=\infty$, P hence Q^* needed for Π^* is (usually) not estimable (from h_∞). On the other hand, for each $\phi\in\Phi$, $p=p_\phi$ can be determined (exactly) by (37) (under weak conditions). From p_ϕ we can determine q^*_ϕ and π^*_ϕ via (8) and (9). The solution always satisfies the reduced Bellman equations exactly, even for very bad reductions, e.g. single state $\phi(h)\equiv 0\,\forall h$. So the reduced problem is not sufficient to judge the quality of ϕ. An alternative to assuming $n=\infty$ is to assume that P is known, which also allows to determine p_ϕ, etc. So what follows applies to stochastic planning as well.

Coarsening and refining reductions ϕ: Let us now coarsen ϕ, i.e. further merge some partitions $\phi^{-1}(s)$. In the simplest case we just merge two states into one. In general, consider coarsening $\chi:\mathcal{S}_\phi\to\mathcal{S}_\psi$ and coarser reduction $\psi:\mathcal{H}\to\mathcal{S}_\psi$ such that $\psi(h)=\chi(\phi(h))$. We also call $\mathcal{S}_\phi$ a refinement of $\mathcal{S}_\psi$. For example, U-trees [McC96, UV98] and Kd-trees have been used in RL [EGW05], where expanding a leaf corresponds to splitting a state. Or in ϕDBN, $S_\phi=\{0,1\}^d$ is a binary feature vector, where removing one component corresponds to pairwise combining 2^d states to 2^{d-1} states [Hut09a].

Ordering reductions in Φ: We can partially order Φ as follows:

$$\begin{aligned}\psi\prec\phi \;&:\Leftrightarrow\; q^*_\phi \text{ and } \pi^*_\phi \text{ are constant on all } s_\phi\in\chi^{-1}(s_\psi) \text{ for all } s_\psi \text{ and } a\\ &\;\Leftrightarrow\; q^*_\phi(s_\phi,a)=q^*_\psi(s_\psi,a) \text{ and } \pi^*_\phi(s_\phi)=\pi^*_\psi(s_\psi) \text{ for all } s_\psi=\chi(s_\phi) \text{ and } a.\end{aligned}$$

$\psi\prec\phi$ means ψ is a better reduction than ϕ since it leads to the same optimal q-value and policy as ψ does, but is more parsimonious (coarser) than ϕ. If q^*_ϕ or π^*_ϕ is not constant on ψ-partitions, coarsening ϕ to ψ and using ψ (potentially) leads to suboptimal solutions.

Enriching the order $\prec$: $\prec$ is a transitive but 'very' partial order. Two maps are incomparable if neither is a refinement of the other. We can enrich order $\prec$ as follows: For any two maps ψ and ψ', the map $\phi(h):=(\psi(h),\psi'(h))\in\mathcal{S}_\phi=\mathcal{S}_\psi\times\mathcal{S}_{\psi'}$ refines both. Define $\psi\prec_\times\psi'$ iff $\psi\prec\phi\prec\psi'$. Extended order $\prec_\times$ is still not total. The remaining incomparable cases are: Case $\psi\prec\phi\succ\psi'$: This is only possible if q^* and π^* of ψ and ψ' (and ϕ) coincide. A secondary criterion based on the relative complexity of ψ and ψ' could decide the case, e.g. $\psi\prec_\times\psi'$ iff $|\mathcal{S}_\psi|<|\mathcal{S}_{\psi'}|$. Case $\psi\succ\phi\prec\psi'$: Both ψ and ψ' are inferior to ϕ. If class Φ is closed under cartesian product, ϕ should be favored over ψ and ψ' so their relative order is not or less important.

Search for ϕ: Assume Φ contains at least one exact reduction. Then the $\prec_\times$-minimal elements in Φ are exactly the maximally coarse exact $\phi\in\Phi$. If Φ is closed under arbitrary coarsening, then there is a unique minimizer (modulo isomorphism). If Φ is also closed under cartesian product, the same holds for $\prec$. This implies that any exhaustive search for a $\prec_\times$-minimum in Φ will give an exact ϕ with minimal number of states, say ϕ_0. Now Theorem 7 tells us that $q^*_{\phi_0}$ and $\pi^*_{\phi_0}$ are the optimal value and policy also of the original process P, irrespective of whether P_{ϕ_0} is Markov or not. So while the conditions of Theorem 7 cannot be verified in practice, the theorem justifies a search procedure based on (q^*_ϕ,π^*_ϕ) that ignores the (non-)Markov structure of P_ϕ.

Search for approximate ϕ based on finite sample size. The principle approach in the previous paragraph is sound, but needs to be generalized in various ways before it can be used: Real sample size is finite, which means we only have access to approximations $\hat{q}^*_\phi$ and $\hat{\pi}^*_\phi$ via estimation $\hat{p}_\phi$ of p_ϕ. The criterion for exact equality $q^*_\phi = q^*_\psi$ in $\prec$ needs to be replaced by a suitable $\hat{q}^*_\phi \approx \hat{q}^*_\psi$, which should be done anyway, since real-word problems seldom allow for exact reductions. $\approx$ should be chosen so as to come with statistical guarantees; e.g. Kolmogorov-Smirnov tests have been used in [McC96]. A suitable $\hat{\pi}^*_\phi \approx \hat{\pi}^*_\psi$ requires more effort (see outlook). For large Φ this also requires appropriate regularization, i.e. penalizing complex ϕ [Hut09c]. To ensure $\hat{q}^* \to q^*$ for $n \to \infty$, we need proper exploration strategies [SLL09]. Finally, we want an efficient search procedure in Φ, rather than exhaustive search. This will be heuristic or will require strong assumptions on Φ [Ngu13]. All but the last point raised above have or should have general solutions (see next paragraph).

Utilizing existing algorithms. The BLB algorithm [MMR11] and its extensions IBLB [NMRO13] and improvements OMS [NOR13] solve most of the problems above and can (nearly) readily be used for our purpose.

The BLB family uses the same basic FRL setup from [Hut09c] used also here. The authors consider a countable class Φ assumed to contain at least one ϕ such that P_ϕ is an MDP (6). They consider average reward, rather then discounting, and analyze regret, which (in general) requires some assumption on the mixing rate or 'diameter' of the MDP. They prove that the total regret grows with $\tilde{O}(n^{1/2...2/3})$, depending on the algorithm.

Their algorithms and analyses rely on UCRL2 [JOA10], an exploration algorithm for finite-state MDPs. Going through the BLB proofs, it appears that the condition that P_ϕ is an MDP can be removed if p (15) is used instead, modulo the analysis of UCRL2 itself. The proofs for the bounds for UCRL2 exploit that s',r' conditioned on s,a are i.i.d., which is true if P_ϕ is Markov but not in general. Asymptotic versions should remain valid under the 'weak conditions' alluded to in (37). With some stronger assumptions that guarantee good convergence rates, the regret analysis of UCRL2 should remain valid too. Formally, the use of Hoeffding's inequality for i.i.d. need to be replaced by comparable bounds with weaker conditions, e.g. Azuma's inequality for martingales.

There is one serious gap in the argument above. BLB uses average reward while our theorems are for discounted reward. It is often possible to adapt algorithms and proofs which come with regret bounds for average reward to PAC bounds for discounted reward or vice versa. This would have to be done first: either a PAC version of BLB by combining MERL [LHS13] with UCRLγ [LH12], or average reward versions of the bounds derived in this paper.

9 Miscellaneous

Action permutation instead of policy condition. We can rename actions without changing the underlying problem: Let $A:\mathcal{A}\to\tilde{\mathcal{A}}$ be a bijection, and define $\tilde{P}(o'r'|h\tilde{a}):=P(o'r'|ha)$, where $\tilde{a}:=A(a)$. Clearly, all results for P also hold for $\tilde{P}$ if a is replaced by $\tilde{a}$ everywhere, in particular $\tilde{\Pi}(h):=A(\Pi(h))$. In general, this is of little use. Things become more interesting if we allow the bijection A to be history-dependent, which we can do since our results hold for any, even non-stationarity, $\tilde{P}$. This allows us to devise an $A:\mathcal{A}\times\mathcal{H}\to\tilde{\mathcal{A}}$ such that $A(\Pi(h);h)=$*constant* for the policy Π of interest. For example, for $\tilde{\mathcal{A}}:=\mathcal{A}$, this is achieved by a permutation that swaps action $a=\Pi(h)$ with some arbitrary but fixed action $a^1\in\mathcal{A}$, and leaves all other actions unchanged:

$$A(a;h) := \begin{cases} a^1 & \text{if} \quad \Pi(h)=a \\ \Pi(h) & \text{if} \quad \Pi(h)\neq a=a^1 \\ a & \text{else} \end{cases}$$

Since $\tilde{\Pi}(h)\equiv A(\Pi(h);h)\equiv a^1$ is constant, the ϕ-uniformity condition for $\tilde{\Pi}$ in Theorems 5, 6 and 9 becomes vacuous. While this transformation is theoretical interest, it only becomes practically useful if we can somehow learn the function A without knowledge of Π, and in particular for Π^*. We could also allow non-bijective A that merge actions that have (approximately) the same (optimal) Q-value.

10 Discussion

Summary. Our results show that RL algorithms for finite-state MDPs can be utilized even for problems P that have arbitrary history dependence and history-to-state reductions/aggregations ϕ that induce P_ϕ that are also neither stationary nor MDP. The only condition to be placed on the reduction is that the quantities of interest, (Q-)Values and (optimal) Policies, can approximately be represented. This considerably generalizes previous work on feature reinforcement learning and MDP state aggregation and allows for extreme state aggregations beyond MDPs. The obtained results may also explain why RL algorithms designed for MDPs sometimes perform well beyond MDPs.

Outlook. As usual, lots remains to be done. A list of the more interesting remaining tasks and open questions follows:

- While the approximate ϕ-uniformity condition on Q^* in Theorem 7 is very weak compared to bisimilarity, uniformity of V^* in Theorem 9 is even weaker (Theorem 11 shows how much of a difference this can make). It is an Open Problem 10 whether an analogue of Theorem 7ii also holds for Theorem 9 beyond $\varepsilon=0$.
- An algorithm learning ϕ beyond MDPs that comes with regret or PAC guarantees has yet to be developed. This could be done by generalizing the partial order

$\prec_\times$ to $n<\infty$, or by adapting the class and proofs of BLB algorithms, or by integrating MERL with UCRLγ. • All bounds contain $\frac{1}{1-\gamma}$ to some power. Can the exponents be improved? For which environments/examples are the bounds tight?

• The trick to use a-dependent Q^* as a-independent map ϕ in Section 6 was to vectorize Q^* in a. Unfortunately this leads to a state-space size exponential in $\mathcal{A}$. Solution ϕ based on (V^*,Π^*) pair is only linear in $\mathcal{A}$, but rests on Open Problem 10. Are there other/better ways of dealing with actions? Other extreme aggregations ϕ, or are a-dependent ϕ possible?

• Are average-reward total-regret versions of our discounted reward results possible, under suitable mixing rate conditions?

• For small discrete action spaces typical for many board games, the exact conditions on Π are met. For continuous action spaces as in robotics, we can simply discretize the action space, introducing another ε-error, but action-continuous versions of our results would be nicer. Except for Theorem 7, any interesting generalization should replace the exact by approximate ϕ-uniformity conditions on Π.

• Our theorems and/or proof ideas should allow to extend existing convergence theorems for RL algorithms such as Q-learning and others from MDPs to beyond MDPs.

• The bisimulation conditions of classical state aggregation results are for reward and transition probabilities. It would be interesting to derive explicit weaker conditions for them that still imply our conditions on (Q-)Values.

References

[EGW05] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[Faz06] I. Fazekas. On a general approach to the strong laws of large numbers. Technical report, Faculty of Informatics, University of Debrecen, Hungary, 2006.

[FK01] I. Fazekas and O. Klesov. A general approach to the strong law of large numbers. *Theory of Probability & Its Applications*, 45(3):436–449, 2001.

[FPP04] N. Ferns, P. Panangaden, and D. Precup. Metrics for finite Markov decision processes. In *Proc. 20th conf. on Uncertainty in Artificial Intelligence (UAI'04)*, pages 162–169, 2004.

[GDG03] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hut09a] M. Hutter. Feature dynamic Bayesian networks. In *Proc. 2nd Conf. on Artificial General Intelligence (AGI'09)*, volume 8, pages 67–73. Atlantis Press, 2009.

[Hut09b] M. Hutter. Feature Markov decision processes. In *Proc. 2nd Conf. on Artificial General Intelligence (AGI'09)*, volume 8, pages 61–66. Atlantis Press, 2009.

[Hut09c] M. Hutter. Feature reinforcement learning: Part I: Unstructured MDPs. *Journal of Artificial General Intelligence*, 1:3–24, 2009.

[JOA10] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[LH12] T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In *Proc. 23rd International Conf. on Algorithmic Learning Theory (ALT'12)*, volume 7568 of *LNAI*, pages 320–334, Lyon, France, 2012. Springer.

[LH14] T. Lattimote and M. Hutter. General time consistent discounting. *Theoretical Computer Science*, 519:140–154, 2014.

[LHS13] T. Lattimore, M. Hutter, and P. Sunehag. The sample-complexity of general reinforcement learning. *Journal of Machine Learning Research, W&CP: ICML*, 28(3):28–36, 2013.

[McC96] A. K. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, University of Rochester, 1996.

[MMR11] O.-A. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS'11)*, volume 24, pages 2627–2635, 2011.

[Ngu13] P. Nguyen. *Feature Reinforcement Learning Agents*. PhD thesis, Research School of Computer Science, Australian National University, 2013.

[NMRO13] P. Nguyen, O. Maillard, D. Ryabko, and R. Ortner. Competing with an infinite set of models in reinforcement learning. *JMLR WS&CP AISTATS*, 31:463–471, 2013.

[NOR13] O.-A. Maillard P. Nguyen, R. Ortner, and D. Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. *JMLR W&CP ICML*, 28(1):543–551, 2013.

[NSH11] P. Nguyen, P. Sunehag, and M. Hutter. Feature reinforcement learning in practice. In *Proc. 9th European Workshop on Reinforcement Learning (EWRL-9)*, volume 7188 of *LNAI*, pages 66–77. Springer, September 2011.

[Ort07] R. Ortner. Pseudometrics for state aggregation in average reward Markov decision processes. In *Proc. 18th International Conf. on Algorithmic Learning Theory (ALT'07)*, volume 4754 of *LNAI*, pages 373–387, Sendai, Japan, 2007.

[Put94] M. L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. Wiley, New York, NY, 1994.

[RN10] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition, 2010.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[SH10] P. Sunehag and M. Hutter. Consistency of feature Markov processes. In *Proc. 21st International Conf. on Algorithmic Learning Theory (ALT'10)*, volume 6331 of *LNAI*, pages 360–374, Canberra, Australia, 2010. Springer.

[SLL09] A. L. Strehl, L. Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

[UV98] W. T. B. Uther and M. M. Veloso. Tree based discretization for continuous state space reinforcement learning. In *AAAI*, pages 769–774, 1998.

[VGS05] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

A List of Notation

General notation

$[\![R]\!]$	$= 1$ if R=true and $=0$ if R=false (Iverson bracket)
$\#\mathcal{X}$	size of set $\mathcal{X}$
ε,δ	small non-negative real numbers
$\lfloor z \rfloor$	largest integer $\leq z$

Original history-based process

$\mathcal{O},\mathcal{R},\mathcal{A}$	$=$ finite observation, reward, action spaces.
$o_t r_t a_t$	$\in \mathcal{O}\times\mathcal{R}\times\mathcal{A}$ = observation, reward, action at time t
$t \leq n \in \mathbb{N}$	$=$ any time $\leq$ sample size
P,Q,V,Π	$=$ Probability, (Q-)Value, Policy of original history-based Process
$\Pi^*,\tilde{\Pi},\Pi_B$	$=$ optimal, approximately optimal, behavior Policy
$h \in \mathcal{H}$	$= (\mathcal{O}\times\mathcal{R}\times\mathcal{A})^*\times\mathcal{O}\times\mathcal{R}$ = possible histories of any length
$h' = hao'r'$	$=$ successor history of $h \in \mathcal{H}$
h_t	$= o_1 r_1 a_1 ... o_t r_t$ = history up to time t
$\mathcal{H}_t$	$= (\mathcal{O}\times\mathcal{R}\times\mathcal{A})^{t-1}\times\mathcal{O}\times\mathcal{R}$ = history of length t
$P(o'r'\|ha)$	$=$ probability of next observation&reward given history&action

Reduction/aggregation from history to states

$\mathcal{S}_\phi$	$=$ finite state space induced by ϕ (range of ϕ)
$\phi:\mathcal{H}\to\mathcal{S}_\phi$	$=$ reduction/map/aggregation from histories to states
s_t	$= \phi(h_t) \in \mathcal{S}$ = state at time t
$P_\phi(s'r'\|ha)$	$=$ marginalized P-probability over state&reward given history&action
$B(h\|sa)$	$=$ dispersion probability. Stochastic "inverse" of ϕ
$\langle Q(h,a)\rangle_B$	$=$ B-average over $\{\tilde{h}:\phi(\tilde{h})=\phi(h)\}$
$w_t(sa)$	$=$ non-negative weight function $\sum_{t=1}^{\infty} w_t(sa) = 1\ \forall sa$
$P_B(h)$	$=$ probability of h from P interacting with Π_B
$P_{\phi B}()$	$=$ (partially) ϕ-reduced, marginalized, conditionalized P_B
$\prec,\prec_\times$	$=$ (extended) ordering of ϕ w.r.t. quality ($n=\infty$ so far only)

Finite state Markov decision process (MDP)

$\mathcal{S}$	$=$ finite state space
p,q,v,π	$=$ probability, (q-)value, policy of MDP
s,a,s',r'	$=$ stat, action, successor state, reward
$n(sas'r')$	$=$ number of times $sas'r'$ appears in h_{n+1}
$\gamma \in [0;1)$	$=$ discount factor