Pseudometrics for State Aggregation in Average Reward Markov Decision Processes

Ronald Ortner

University of Leoben, A-8700 Leoben, Austria
ronald.ortner@unileoben.ac.at

Abstract. We consider how state similarity in average reward Markov decision processes (MDPs) may be described by pseudometrics. Introducing the notion of *adequate* pseudometrics which are well adapted to the structure of the MDP, we show how these may be used for state aggregation. Upper bounds on the loss that may be caused by working on the aggregated instead of the original MDP are given and compared to the bounds that have been achieved for discounted reward MDPs.

1 Introduction

Most work done in hierarchical reinforcement learning, relational reinforcement learning, function approximation, factorization and state aggregation ultimately addresses the problem of how to deal with large state spaces in Markov decision processes (MDPs). Here we are concerned with *state aggregation* (for references see [1]), which tries to convert the idea that similar states (with respect to rewards and transition probabilities) may be aggregated to meta-states, and calculation of the optimal policy may then be conducted on the meta-MDP.

For discounted reward MDPs, upper bounds on the loss that may be caused by aggregation have been obtained by Even-Dar and Mansour [2] and more recently by Ferns et al. [3]. We are particularly interested in the latter work, as it has introduced the idea that state similarity may be described by pseudometrics. Here we try to extend this approach, first by giving a general definition of *adequate* metrics which are useful for state aggregation, and secondly by generalizing the results of [3] and [2] to average reward MDPs.

The paper is organized as follows. After preliminary definitions in Sect. 2, we show in Sect. 3 how to conduct state aggregation with respect to a given metric. We consider a very simple distance function d_v and give an upper bound on the loss by state aggregation with respect to d_v. Then in Sect. 4, we generally define *adequate* distance functions and generalize the results accordingly. In Sect. 5, we compare our bounds to those obtained in the discounted case and show why the loss by aggregation may be significantly larger for average reward MDPs. In the final section, we consider basic questions on the possibility of online aggregation and other open problems for future research.

2 Preliminaries

Definition 1. *A* Markov decision process (MDP) $\mathcal{M} = \langle S, A, \mu_0, p, r\rangle$ *consists of* ***(i)*** *a finite set of* states S *with* ***(ii)*** *a finite set of* actions A *available in each state* $\in S$, ***(iii)*** *an initial distribution* μ_0 *over* S, ***(iv)*** *the transition probabilities* $p_a(s, s')$ *which give the probability of reaching state* s' *when choosing action* a *in state* s, *and* ***(v)*** *the payoff distributions with mean* $r_a(s)$ *and support in* $[0, 1]$ *that specify the random reward obtained for choosing action* a *in state* s.

A *policy* on an MDP $\mathcal{M}$ is a mapping $\pi : S \to A$. Note that each policy π induces a Markov chain $\mathcal{M}_\pi$ on $\mathcal{M}$. We will only consider *ergodic* MDPs, where all policies induce ergodic Markov chains (in which states are reachable from each other after a finite number of steps). For a policy π let μ_π be the *stationary distribution* of $\mathcal{M}_\pi$. Remember that for ergodic Markov chains with probability matrix P this is the unique distribution μ with $\mu P = \mu$ (cf. e.g. [4]). The *average reward of* π then may be defined as

$$\rho_\pi(\mathcal{M}) := \sum_{s \in S} \mu_\pi(s)\, r_{\pi(s)}(s).$$

A policy π^* is *optimal on* $\mathcal{M}$, if $\rho_\pi(\mathcal{M}) \leq \rho_{\pi^*}(\mathcal{M}) =: \rho^*$ for all policies π. As ρ_π is independent of the initial distribution μ_0, in the following we ignore μ_0 and write MDPs as tuples $\mathcal{M} = \langle S, A, p, r\rangle$.

Definition 2. *Given a set* X *and a nonnegative function* $d : X \times X \to \mathbb{R}$, *we call* (X, d) *a* pseudometric space *with* pseudometric d, *if for all* $x, y, z \in X$,

$$\begin{aligned} &(i) \quad d(x, x) = 0, \\ &(ii) \quad d(x, y) = d(y, x), \\ &(iii) \quad d(x, y) + d(y, z) \leq d(x, z). \end{aligned}$$

In general, for d being a *metric on* X it is additionally demanded that $d(x, y) = 0$ implies $x = y$. As we will consider pseudometrics on state spaces of MDPs, this is obviously not a desired property (i.e., we want to include the possibility of having distinct states with equal properties).

Definition 3. *Given a Markov chain* $\mathcal{C}$ *with state space* S *and stationary distribution* μ, *its* mixing time with respect to state s *is defined as*

$$\kappa_s := \sum_{s' \in S} m_{ss'} \mu(s'),$$

where $m_{ss'}$ *is the* mean first passage time from s to s' *if* $s \neq s'$, *while* m_{ss} *is the* mean return time to s. *It can be shown that* κ_s *is independent of* s *(see [5]), so that may speak of the the* mixing time of $\mathcal{C}$, *denoted by* $\kappa_\mathcal{C}$.

3 A Simple Pseudometric for State Similarity

3.1 Block MDPs

Definition 4. *An MDP $\mathcal{M} = \langle S, A, p, r\rangle$ is a* block MDP *with blocks $S_1, \ldots, S_k$, if the block set $\{S_1, \ldots, S_k\}$ is a partition of S, and for all $a \in A$, all $s'' \in S$, and all s, s' in the same block S_i,*

$$r_a(s) = r_a(s'), \text{ and } \quad p_a(s, s'') = p_a(s', s'').$$

A policy π on a block MDP is called uniform, *if $\pi(s) = \pi(s')$ for s, s' in the same block.*

Obviously, block MDPs are predestined to be aggregated. However, the following definition is also applicable to arbitrary MDPs.

Definition 5. *Given an MDP $\mathcal{M} = \langle S, A, p, r\rangle$ and a partition $\widehat{S} = \{S_1, \ldots, S_k\}$ of its state space S, the* aggregated MDP with respect to $\widehat{S}$ *is defined as $\widehat{\mathcal{M}} := \langle \widehat{S}, A, \widehat{p}, \widehat{r}\rangle$, where*

$$\widehat{r}_a(S_i) := \frac{1}{|S_i|} \sum_{s \in S_i} r_a(s), \text{ and } \quad \widehat{p}_a(S_i, S_j) := \frac{1}{|S_i|} \sum_{s \in S_i} \sum_{s' \in S_j} p_a(s, s').$$

It is easy to check that $\widehat{p}_a(S_i, \cdot)$ is a probability distribution for each $S_i \in \widehat{S}$.

Any policy π on an aggregated MDP $\widehat{\mathcal{M}}$ with state space $\widehat{S} = \{S_1, \ldots, S_k\}$ can be naturally extended to a policy π^e on the original MDP $\mathcal{M}$ by

$$\pi^e(s) := a, \text{ if } s \in S_j \text{ and } \pi(S_j) = a.$$

We continue with some considerations on block MDPs, the first one being trivial if the stationary distribution μ in state s is interpreted as probability of being in s after an infinite number of steps. However, we give a proof which refers only to the properties of stationary distributions mentioned in Sect. 2.

Lemma 1. *Let $\mathcal{M} = \langle S, A, p, r\rangle$ be a block MDP with block set $\widehat{S} = \{S_1, \ldots, S_k\}$ and respective aggregated MDP $\widehat{\mathcal{M}} = \langle \widehat{S}, A, \widehat{p}, \widehat{r}\rangle$. Given a policy π on $\widehat{\mathcal{M}}$ and its extended counterpart π^e on $\mathcal{M}$ with stationary distributions $\widehat{\mu}_\pi$ and μ_{π^e}, respectively, for all $S_i \in \widehat{S}$,*

$$\widehat{\mu}_\pi(S_i) = \sum_{s \in S_i} \mu_{\pi^e}(s).$$

Proof. First, note that since $\mathcal{M}$ is a block MDP, for all $s \in S_j$ and $a \in A$,

$$\widehat{p}_a(S_j, S_i) = \sum_{s' \in S_i} p_a(s, s'). \tag{1}$$

As $\mu P = \mu$ for the stationary distribution μ of a transition matrix P, we have[1] for all $s' \in S$,

$$\sum_{s \in S} \mu_{\pi^e}(s)\, p(s, s') = \mu_{\pi^e}(s'). \tag{2}$$

Let $\widehat{P}$ be the transition matrix of $\widehat{\mathcal{M}}$ under π. We set $\mu'(S_j) := \sum_{s \in S_j} \mu_{\pi^e}(s)$ for $S_j \in \widehat{S}$, and have by (2) and (1) for each $S_i \in \widehat{S}$,

$$\begin{aligned}
(\mu' \widehat{P})_{S_i} &= \sum_{S_j \in \widehat{S}} \mu'(S_j)\, \widehat{p}(S_j, S_i) \;=\; \sum_{S_j \in \widehat{S}} \sum_{s \in S_j} \mu_{\pi^e}(s)\; \widehat{p}(S_j, S_i) \\
&= \sum_{S_j \in \widehat{S}} \sum_{s \in S_j} \mu_{\pi^e}(s) \sum_{s' \in S_i} p(s, s') \;=\; \sum_{s' \in S_i} \sum_{S_j \in \widehat{S}} \sum_{s \in S_j} \mu_{\pi^e}(s)\; p(s, s') \\
&= \sum_{s' \in S_i} \sum_{s \in S} \mu_{\pi^e}(s)\; p(s, s') \;=\; \sum_{s' \in S_i} \mu_{\pi^e}(s') \;=\; \mu'(S_i).
\end{aligned}$$

Consequently, by the uniqueness of the stationary distribution we have $\widehat{\mu}_\pi = \mu'$, which proves the lemma. □

Theorem 1. *Each block MDP has an optimal policy which is uniform.*

In the proof of Theorem 1 we will make use of a minor result about optimal policies on ergodic MDPs.

Definition 6. *Given policies $\pi_1, \ldots, \pi_\ell$ on an MDP with state space S, a policy π is called a* combination of $\pi_1, \ldots, \pi_\ell$*, if for each $s \in S$ there is an $i \in \{1, \ldots, \ell\}$ such that $\pi(s) = \pi_i(s)$.*

The following proposition can be derived from the Bellman equations, which may also be used to prove Theorem 1 directly (cf. the proof of the more general Theorem 4 in Sect. 4 below). As a corollary to a more general result Proposition 1 has been proved in [6].

Proposition 1. *On ergodic MDPs, any combination of optimal policies is optimal.*

Proof of Theorem 1. Consider an arbitrary non-uniform, optimal policy π^* on a block MDP $\mathcal{M}$ with blocks $S_1, \ldots, S_k$. Take some block $S_j = \{s_1, \ldots, s_m\}$ on which π^* is not uniform. As $\mathcal{M}$ is a block MDP, all states in S_j have the same rewards and transition probabilities under each action $a \in A$. Hence, a policy π is optimal, if it coincides with π^* on $S \setminus S_j$ and swaps the actions in S_j according to some permutation $\sigma : S_j \to S_j$, that is, $\pi(s_i) = \pi^*\big(\sigma(s_i)\big)$ for $i = 1, \ldots, m$.

Thus in particular, for each $i \in \{1, \ldots, m\}$ there is an optimal policy π such that $\pi(s_i) = \pi^*(s_1)$. It follows from Proposition 1 that there is an optimal policy which is uniform on S_j. This argument can be repeated for each single block to yield the theorem. □

[1] In the following, we usually skip indices for actions when the policy is fixed.

3.2 A Simple Pseudometric, ε-Aggregation, and an Upper Bound

Given an MDP $\mathcal{M} = \langle S, A, p, r\rangle$ and positive constants c_r, c_p, we set for $s, s' \in S$,

$$d_v(s, s') := \max_{a\in A} \Big\{ c_r \left| r_a(s) - r_a(s') \right| + c_p \sum_{s''\in S} \left| p_a(s, s'') - p_a(s', s'') \right| \Big\}.$$

It is easy to check that d_v is a pseudometric on S. However, d_v is not a metric. If $d_v(s, s') = 0$, then all rewards and transition probabilities coincide in states s and s', which however does not entail that $s = s'$. The pseudometric d_v is basically the *bisimulation metric* induced by the *total variation probability metric*, which has been introduced for discounted MDPs in [3]. Ferns et al. consider also other probability metrics that measure the distance between two transition probability distributions $p_a(s, \cdot)$ and $p_a(s', \cdot)$.

Definition 7. *For fixed $\varepsilon > 0$, an* ε-partition of the state space S with respect to a pseudometric d *on S is a minimal partition of S into* aggregated states *(or* blocks*) $S_1, \ldots, S_k$ such that for $s, s' \in S_i$ one has $d(s, s') < \varepsilon$. Minimality here means that one cannot aggregate any S_i, S_j to $S_i \cup S_j$, that is, for distinct S_i, S_j there are $s \in S_i$, $s' \in S_j$ with $d(s, s') \geq \varepsilon$.*

When aggregating an MDP $\mathcal{M}$ with respect to an ε-partition we speak of an ε-aggregation of $\mathcal{M}$.

Theorem 2. *Let $\mathcal{M} = \langle S, A, p, r\rangle$ be an MDP and $\widehat{\mathcal{M}} = \langle \widehat{S}, A, \widehat{p}, \widehat{r}\rangle$ an ε-aggregation of $\mathcal{M}$ with respect to d_v. Then for each policy π on $\widehat{\mathcal{M}}$ and its respective extended policy π^e on $\mathcal{M}$,*

$$\left| \rho_{\pi^e}(\mathcal{M}) - \rho_\pi(\widehat{\mathcal{M}}) \right| \;<\; \left(\frac{1}{c_r} + \frac{\kappa_{\mathcal{M}_\pi} - 1}{c_p} \right) \varepsilon,$$

where $\kappa_{\mathcal{M}_\pi}$ is the mixing time of the Markov chain induced by π on $\mathcal{M}$.

For the proof of Theorem 2 we will need the following result of [5] on perturbations of Markov chains.

Theorem 3 (Hunter[5]). *Let $\mathcal{C}$, $\widetilde{\mathcal{C}}$ be two ergodic Markov chains on the same state space S with transition probabilities $p(\cdot, \cdot)$, $\widetilde{p}(\cdot, \cdot)$ and stationary distributions μ, $\widetilde{\mu}$. Then*

$$\left\| \mu - \widetilde{\mu} \right\|_1 \;\leq\; (\kappa_{\mathcal{C}} - 1) \max_{s\in S} \sum_{s'\in S} \left| p(s, s') - \widetilde{p}(s, s') \right|,$$

where $\kappa_{\mathcal{C}}$ is the mixing time of $\mathcal{C}$.

Proof of Theorem 2. Let us first modify the original MDP $\mathcal{M}$ by redefining the rewards in each state $s \in S_j$ $(1 \leq j \leq k := |\widehat{S}|)$ and each $a \in A$ as

$$\widetilde{r}_a(s) := \frac{1}{|S_j|} \sum_{s'\in S_j} r_a(s').$$

Then using the assumption that two states s, s' in the same block S_j have distance $d_v(s,s') < \varepsilon$, the difference in the average rewards of the original and the thus modified MDP $\mathcal{M}_{\widetilde{r}} := \langle S, A, p, \widetilde{r}\rangle$ under some fixed policy π can be upper bounded by

$$\begin{aligned}
&\big|\rho_\pi(\mathcal{M}) - \rho_\pi(\mathcal{M}_{\widetilde{r}})\big| = \\
&= \Big|\sum_{s\in S}\mu_\pi(s)\,r(s) - \sum_{s\in S}\mu_\pi(s)\,\widetilde{r}(s)\Big| \;=\; \Big|\sum_{j=1}^{k}\sum_{s\in S_j}\mu_\pi(s)\Big(r(s) - \frac{1}{|S_j|}\sum_{s'\in S_j} r(s')\Big)\Big| \\
&= \Big|\sum_{j=1}^{k}\sum_{s\in S_j}\mu_\pi(s)\Big(\frac{1}{|S_j|}\sum_{s'\in S_j}\big(r(s) - r(s')\big)\Big)\Big| \;<\; \sum_{j=1}^{k}\sum_{s\in S_j}\mu_\pi(s)\Big(\frac{1}{|S_j|}\sum_{s'\in S_j}\frac{\varepsilon}{c_r}\Big) \\
&= \sum_{s\in S}\mu_\pi(s)\frac{\varepsilon}{c_r} \;=\; \frac{\varepsilon}{c_r}. \qquad (3)
\end{aligned}$$

Now we also redefine the transition probabilities for $s \in S_j$ and $a \in A$ to be

$$\widetilde{p}_a(s,s') := \frac{1}{|S_j|}\sum_{s''\in S_j} p_a(s'',s').$$

It is easily checked that the $p_a(s,\cdot)$ are indeed probability distributions for all $s \in S$. Considering any policy π, for each $s \in S_j$,

$$\begin{aligned}
&\sum_{s'\in S}\big|p(s,s') - \widetilde{p}(s,s')\big| \;=\; \sum_{s'\in S}\Big|p(s,s') - \frac{1}{|S_j|}\sum_{s''\in S_j} p(s'',s')\Big| \\
&= \sum_{s'\in S}\Big|\frac{1}{|S_j|}\sum_{s''\in S_j}\big(p(s,s') - p(s'',s')\big)\Big| \;\le\; \sum_{s'\in S}\frac{1}{|S_j|}\sum_{s''\in S_j}\big|p(s,s') - p(s'',s')\big| \\
&= \frac{1}{|S_j|}\sum_{s''\in S_j}\sum_{s'\in S}\big|p(s,s') - p(s'',s')\big| \;<\; \frac{1}{|S_j|}\sum_{s''\in S_j}\frac{\varepsilon}{c_p} \;=\; \frac{\varepsilon}{c_p}, \qquad (4)
\end{aligned}$$

again using that states s, s'' in the same block have distance $d_v(s,s'') < \varepsilon$. As rewards are upper bounded by 1, Theorem 3 and (4) give for the difference of the average rewards of $\mathcal{M}_{\widetilde{r}}$ and $\widetilde{\mathcal{M}} := \langle S, A, \widetilde{p}, \widetilde{r}\rangle$ under policy π (with respective stationary distributions μ_π and $\widetilde{\mu}_\pi$),

$$\begin{aligned}
\big|\rho_\pi(\mathcal{M}_{\widetilde{r}}) - \rho_\pi(\widetilde{\mathcal{M}})\big| &= \Big|\sum_{s\in S}\mu_\pi(s)\,\widetilde{r}(s) - \sum_{s\in S}\widetilde{\mu}_\pi(s)\,\widetilde{r}(s)\Big| = \\
&= \Big|\sum_{s\in S}\big(\mu_\pi(s) - \widetilde{\mu}_\pi(s)\big)\,\widetilde{r}(s)\Big| \;\le\; \sum_{s\in S}\big|\mu_\pi(s) - \widetilde{\mu}_\pi(s)\big|\,\widetilde{r}(s) \\
&\le \sum_{s\in S}\big|\mu_\pi(s) - \widetilde{\mu}_\pi(s)\big| \;=\; \big\|\mu_\pi - \widetilde{\mu}_\pi\big\|_1 \;<\; (\kappa_{\mathcal{M}_\pi} - 1)\frac{\varepsilon}{c_p}.
\end{aligned}$$

Combining this with (3) yields

$$\begin{aligned}\left|\rho_\pi(\mathcal{M}) - \rho_\pi(\widetilde{\mathcal{M}})\right| &\leq \left|\rho_\pi(\mathcal{M}) - \rho_\pi(\mathcal{M}_{\widetilde{r}})\right| + \left|\rho_\pi(\mathcal{M}_{\widetilde{r}}) - \rho_\pi(\widetilde{\mathcal{M}})\right| \\ &< \frac{\varepsilon}{c_r} + (\kappa_{\mathcal{M}_\pi} - 1)\frac{\varepsilon}{c_p}. \end{aligned} \tag{5}$$

So far, π has been an arbitrary policy on $\mathcal{M}$. Now we fix π to be a policy on $\widehat{\mathcal{M}}$ and claim that $\rho_{\pi^e}(\widetilde{\mathcal{M}}) = \rho_\pi(\widehat{\mathcal{M}})$ for the extension π^e of π. It is easy to see that by definition of the rewards and transition probabilities, $\widetilde{\mathcal{M}}$ is a block MDP with block set $\widehat{S}$ and respective aggregated MDP $\widehat{\mathcal{M}}$. In particular, $\widehat{r}_a(S_j) = \widetilde{r}_a(s)$ for all $a \in A$ and $s \in S_j$, so that by Lemma 1

$$\rho_\pi(\widehat{\mathcal{M}}) = \sum_{S_j \in \widehat{S}} \widehat{\mu}_\pi(S_j)\,\widehat{r}(S_j) = \sum_{S_j \in \widehat{S}} \sum_{s \in S_j} \widetilde{\mu}_{\pi^e}(s)\,\widetilde{r}(s) = \sum_{s \in S} \widetilde{\mu}_{\pi^e}(s)\,\widetilde{r}(s) = \rho_{\pi^e}(\widetilde{\mathcal{M}}),$$

which together with (5) proves the theorem. □

Corollary 1. *Let π^* be an optimal policy on an MDP $\mathcal{M}$ with optimal average reward $\rho^* := \rho_{\pi^*}(\mathcal{M})$, and let $\widehat{\pi}^*$ be an optimal policy with optimal average reward $\widehat{\rho}^* := \rho_{\widehat{\pi}^*}(\widehat{\mathcal{M}})$ on an ε-aggregation $\widehat{\mathcal{M}}$ of $\mathcal{M}$ with respect to d_v. Then*

$$\begin{aligned} &\textit{(i)} \quad |\rho^* - \widehat{\rho}^*| < \left(\frac{1}{c_r} + \frac{\kappa_{\mathcal{M}} - 1}{c_p}\right)\varepsilon, \\ &\textit{(ii)} \quad \rho^* < \rho_{\widehat{\pi}^{*e}}(\mathcal{M}) + \left(\frac{2}{c_r} + \frac{2(\kappa_{\mathcal{M}} - 1)}{c_p}\right)\varepsilon, \end{aligned}$$

where $\kappa_{\mathcal{M}} := \max_\pi \kappa_{\mathcal{M}_\pi}$.

Proof. First note that the extension $\widehat{\pi}^{*e}$ of $\widehat{\pi}^*$ to the block MDP $\widetilde{\mathcal{M}}$ (as defined in the proof of Theorem 2) is optimal on $\widetilde{\mathcal{M}}$ with reward $\widehat{\rho}^*$. This follows from Theorem 1 and the fact that $\rho_{\pi^e}(\widetilde{\mathcal{M}}) = \rho_\pi(\widehat{\mathcal{M}})$ (cf. proof of Theorem 2). Now if $\rho^* > \widehat{\rho}^*$, then by optimality of $\widehat{\pi}^{*e}$ on $\widetilde{\mathcal{M}}$,

$$\rho_{\pi^*}(\mathcal{M}) \;=\; \rho^* \;>\; \widehat{\rho}^* \;=\; \rho_{\widehat{\pi}^{*e}}(\widetilde{\mathcal{M}}) \;\geq\; \rho_{\pi^*}(\widetilde{\mathcal{M}}),$$

so that by (5),

$$|\rho^* - \widehat{\rho}^*| \;\leq\; \left|\rho_{\pi^*}(\mathcal{M}) - \rho_{\pi^*}(\widetilde{\mathcal{M}})\right| \;<\; \frac{\varepsilon}{c_r} + (\kappa_{\pi^*} - 1)\frac{\varepsilon}{c_p}. \tag{6}$$

On the other hand, if $\rho^* \leq \widehat{\rho}^*$, then by optimality of π^* on $\mathcal{M}$,

$$\rho_{\widehat{\pi}^{*e}}(\widetilde{\mathcal{M}}) \;=\; \widehat{\rho}^* \;\geq\; \rho^* \;=\; \rho_{\pi^*}(\mathcal{M}) \;\geq\; \rho_{\widehat{\pi}^{*e}}(\mathcal{M}),$$

and it follows again from (5) that

$$|\widehat{\rho}^* - \rho^*| \;\leq\; \left|\rho_{\widehat{\pi}^{*e}}(\widetilde{\mathcal{M}}) - \rho_{\widehat{\pi}^{*e}}(\mathcal{M})\right| \;<\; \frac{\varepsilon}{c_r} + (\kappa_{\widehat{\pi}^{*e}} - 1)\frac{\varepsilon}{c_p},$$

which together with (6) finishes the proof of (i).

Concerning (ii), note that by optimality of $\widehat{\pi}^{*e}$ on $\widetilde{\mathcal{M}}$ it follows from (5) that

$$\begin{aligned}
\rho^* - \rho_{\widehat{\pi}^{*e}}(\mathcal{M}) &\leq \rho^* - \rho_{\widehat{\pi}^{*e}}(\mathcal{M}) + \big(\widehat{\rho}^* - \rho_{\pi^*}(\widetilde{\mathcal{M}})\big) \\
&= \rho_{\pi^*}(\mathcal{M}) - \rho_{\pi^*}(\widetilde{\mathcal{M}}) + \rho_{\widehat{\pi}^{*e}}(\widetilde{\mathcal{M}}) - \rho_{\widehat{\pi}^{*e}}(\mathcal{M}) \\
&\leq \big|\rho_{\pi^*}(\mathcal{M}) - \rho_{\pi^*}(\widetilde{\mathcal{M}})\big| + \big|\rho_{\widehat{\pi}^{*e}}(\widetilde{\mathcal{M}}) - \rho_{\widehat{\pi}^{*e}}(\mathcal{M})\big| \\
&< 2\bigg(\frac{\varepsilon}{c_r} + (\kappa_{\mathcal{M}} - 1)\frac{\varepsilon}{c_p}\bigg). \qquad \square
\end{aligned}$$

Theorem 2 and Corollary 1 (i) can be seen as generalizations of the bounds for discounted reward MDPs obtained in Theorem 5.2 of [3].

4 Adequate Similarity Metrics

Obviously, ε-aggregation with respect to d_v is a rather restricted model which will be applicable only to very special problems. In this section, we want to develop a more general view on similarity metrics on an MDP's state space.

4.1 Generalized Block MDPs

Definition 8. *An MDP $\mathcal{M} = \langle S, A, p, r\rangle$ is a* generalized block MDP *with blocks $S_1, \ldots, S_k$, if the block set $\{S_1, \ldots, S_k\}$ is a partition of S, and for all s, s' in the same block S_i, all $a \in A$, and all blocks S_j there is an $a' \in A$ such that*

$$r_a(s) = r_{a'}(s'), \text{ and } \quad \sum_{s'' \in S_j} p_a(s, s'') = \sum_{s'' \in S_j} p_{a'}(s', s''). \tag{7}$$

With this definition, we could also consider MDPs in which each state has an individual set of possible actions at its disposal. All results presented easily generalize to this setting. However, for the sake of simplicity, we assume in the following without loss of generality that within a block S_i the actions in A are labelled uniformly, such that for states $s, s' \in S_i$, (7) holds for $a' = a$. Consequently, we may define *uniform policies* as we have done before.

Generalized block MDPs (yet with discounted rewards) have already been considered by Givan et al. [1] under the name of *stochastic bisimulation*, which is the equivalence relation that corresponds to the partition $\{S_1, \ldots, S_k\}$ in Definition 8 (cf. also the discussion in [3]).

Note that block MDPs are also generalized block MDPs, so that most results in this section can be considered as generalizations of the results in the previous section.

Lemma 2. *Let $\mathcal{M} = \langle S, A, p, r\rangle$ be a generalized block MDP with block set $\widehat{S} = \{S_1, \ldots, S_k\}$ and respective aggregated MDP $\widehat{\mathcal{M}} = \langle \widehat{S}, A, \widehat{p}, \widehat{r}\rangle$. Given a policy π*

on $\widehat{\mathcal{M}}$ and its extended counterpart π^e on $\mathcal{M}$ with stationary distributions $\widehat{\mu}_\pi$ and μ_{π^e}, respectively, one has for all $S_j \in \widehat{S}$,

$$\widehat{\mu}_\pi(S_j) = \sum_{s \in S_j} \mu_{\pi^e}(s).$$

Proof. As proof of Lemma 1. □

Theorem 4. *On generalized block MDPs there is always a uniform policy which gives optimal average return.*

Proof. Let $\mathcal{M} = \langle S, A, p, r \rangle$ be a generalized block MDP with block set $\widehat{S} = \{S_1, \ldots, S_k\}$. It is a well-known fact (cf. e.g. [7]) that a policy on an ergodic MDP is optimal if it solves the Bellman equations, that is, if there is ρ^* and a value function $v : S \to \mathbb{R}$ such that for all $s \in S$,

$$v(s) + \rho^* = \max_{a \in A} \Big(r_a(s) + \sum_{s' \in S} p_a(s, s')\, v(s') \Big). \tag{8}$$

Thus, an optimal policy $\widehat{\pi}^*$ on the aggregated MDP $\widehat{\mathcal{M}} = \langle \widehat{S}, A, \widehat{p}, \widehat{r} \rangle$ solves for all $S_i \in \widehat{S}$,

$$\begin{aligned} \widehat{v}(S_i) + \widehat{\rho}^* &= \max_{a \in A} \Big(\widehat{r}_a(S_i) + \sum_{S_j \in \widehat{S}} \widehat{p}_a(S_i, S_j)\, \widehat{v}(S_j) \Big) \\ &= \widehat{r}_{\widehat{\pi}^*(S_i)}(S_i) + \sum_{S_j \in \widehat{S}} \widehat{p}_{\widehat{\pi}^*(S_i)}(S_i, S_j)\, \widehat{v}(S_j). \end{aligned} \tag{9}$$

However, setting $v(s) := v(S_j)$ for $s \in S_j$, it follows from (9) that the Bellman equations (8) hold for the extension $\widehat{\pi}^{*e}$ of $\widehat{\pi}^*$ to $\mathcal{M}$ for all $s \in S$, which means that $\widehat{\pi}^{*e}$ is optimal on $\mathcal{M}$. □

4.2 Adequate Similarity Metrics

The key idea an *adequate* similarity metric shall grasp is that in similar states there should be equivalent actions available which lead to similar states again. Such a metric may then be used to partition the state space. As similarity in general is not a transitive relation, not any partition will work (for more about the problem of obtaining adequate partitions from similarity relations see e.g. [8]). Thus before formalizing our basic idea, we start with a condition for the utility of a given partition induced by a distance metric.

Definition 9. *Given $\varepsilon > 0$ and a pseudometric space (S, d), we say that $S' \subseteq S$ is* ε-maximal*, if (i) for all $s, s' \in S'$, $d(s, s') < \varepsilon$, and (ii) for all $s'' \in S \setminus S'$ there is $s \in S'$ with $d(s, s'') \geq \varepsilon$.*

An ε-partition $\widehat{S} = \{S_1, \ldots, S_k\}$ of S with respect to a metric d is called consistent*, if each $S_i \in \widehat{S}$ is ε-maximal.*

Unfortunately, existence of consistent ε-aggregations of the state space cannot be guaranteed for each $\varepsilon > 0$.

Example 1. Let $S = \{s_1, s_2, s_3\}$ with $d(s_1, s_2), d(s_2, s_3) < \varepsilon$ and $d(s_1, s_3) \geq \varepsilon$. Then neither of the two possible ε-partitions $\widehat{S}_1 = \big\{\{s_1, s_2\}, \{s_3\}\big\}$ and $\widehat{S}_2 = \big\{\{s_1\}, \{s_2, s_3\}\big\}$ is consistent, because the singletons $\{s_3\}$ and $\{s_1\}$ are not ε-maximal.

Sometimes, things are easier if (S, d) can be embedded into some larger metric space (X, d), e.g. if $S \subset \mathbb{R}^n$ and d coincides on S with some arbitrary metric d on $\mathbb{R}^n$. In this case one may relax the condition for ε-maximality as follows:

A set $S' \subseteq S$ is *ε-maximal*, if $S' = S \cap U_\varepsilon(x)$ for some ε-ball $U_\varepsilon(x) := \{y \in X : d(x, y) < \varepsilon\}$ with center $x \in X$. Then an ε-partition $\widehat{S} = \{S_1, \ldots, S_k\}$ is *consistent* if it can be represented by non-intersecting ε-balls, that is, if

(i) there are $x_1, \ldots, x_k \in \mathbb{R}^n$ such that $S_i = S \cap U_\varepsilon(x_i)$ for $i = 1, \ldots, k$,
(ii) $U_\varepsilon(x_i) \cap U_\varepsilon(x_j) = \varnothing$ for $i \neq j$.

However, such an embedding may fail to give consistency either.

Example 2. Let $S = \{s_1, s_2, s_3\}$ consist of three points s_1, s_2, s_3 equidistantly distributed on a circle $C := \{y \in \mathbb{R}^2 : \|x - y\|_2 = r\}$ with center x and radius r. Considering the metric space (C, d) with $d(y, z) := \|y - z\|_2$ for $y, z \in C$, it is easy to see that for $\varepsilon = \sqrt{2}r$ (so that for $y \in C$, $U_\varepsilon(y)$ contains one half of C), there is no consistent ε-partition of S. This example can easily be extended to arbitrary n-dimensional spheres.

Also, $\mathbb{R}^n$ with Euclidean distance may not be favorable anyway, as it is impossible to cover $\mathbb{R}^n$ with non-intersecting ε-balls with respect to Euclidean distance. Thus, the metric with respect to $\|\cdot\|_\infty$, which evidently guarantees a consistent ε-partition in $\mathbb{R}^n$ for each $\varepsilon > 0$, will be preferred.

Definition 10. *Given an MDP $\mathcal{M} = \langle S, A, p, r\rangle$, we say that a pseudometric d on S is* adequate *to $\mathcal{M}$, if $d(s, s') < \varepsilon$ implies that for all $a \in A$ there is an $a' \in A$ such that*

$$\textit{(i)}\quad c_r \left| r_a(s) - r_{a'}(s') \right| < \varepsilon,$$
$$\textit{(ii)}\quad c_p \Big| \sum_{s'' \in S'} p_a(s, s'') - \sum_{s'' \in S'} p_{a'}(s', s'') \Big| < \varepsilon \quad \textit{for all } \varepsilon\textit{-maximal } S' \subseteq S.$$

As in the case of generalized block MDPs we assume without loss of generality that for states s, s' in the same block, actions are labelled uniformly so that $a' = a$ in the definition above.

Of course, one may as well define a particular partition $\widehat{S} = \{S_1, \ldots, S_k\}$ of the state space to be *ε-adequate*, if for all s, s' in the same block S_j,

$$\text{(i)}\quad c_r \left| r_a(s) - r_a(s') \right| < \varepsilon,$$
$$\text{(ii')}\quad c_p \Big| \sum_{s'' \in S_i} p_a(s, s'') - \sum_{s'' \in S_i} p_a(s', s'') \Big| < \varepsilon \quad \text{for all } S_i \in \widehat{S}.$$

This modified definition is similar to the definition of ε-*homogeneous* partitions for discounted reward MDPs in [2]. The only difference is that in condition (ii'), Even-Dar and Mansour consider arbitrary norms and sum up over all aggregated states.

Further, one still may work with the metric d_v defined in the previous section. Even though the kind of state similarity which may be grasped by d_v is rather restricted, aggregating states with respect to d_v for given $\varepsilon > 0$ evidently gives ε-adequate partitions of the state space. By definition, d_v is also an adequate metric.

4.3 A General Upper Bound on the Loss by Aggregation

By the remarks at the end of the previous section, the following theorem can be seen as a generalization of Lemma 3 of [2] to average reward MDPs. More importantly, for average reward MDPs a similar result has been given by Ren and Krogh [9].[2] Our theorem is however an improvement, as the used mixing time parameter is smaller than the respective parameter in [9].

Theorem 5. *Given an MDP $\mathcal{M} = \langle S, A, p, r\rangle$ and a consistent ε-aggregation $\widehat{\mathcal{M}} = \langle \widehat{S}, A, \widehat{p}, \widehat{r}\rangle$ of $\mathcal{M}$ with respect to an adequate pseudometric d, for each policy π on $\widehat{\mathcal{M}}$ and its respective extended policy π^e on $\mathcal{M}$,*

$$\left|\rho_{\pi^e}(\mathcal{M}) - \rho_\pi(\widehat{\mathcal{M}})\right| \;<\; \left(\frac{1}{c_r} + \frac{(\kappa_{\mathcal{M}_\pi} - 1)|\widehat{S}|}{c_p}\right)\varepsilon.$$

Proof. As in the proof of Theorem 2, we start by modifying the rewards in $\mathcal{M}$ slightly to be

$$\widetilde{r}_a(s) := \frac{1}{|S_j|}\sum_{s'\in S_j} r_a(s') \tag{10}$$

for $s \in S_j$ and $a \in A$. Then the same argument can be repeated to see that for the modified MDP $\mathcal{M}_{\widetilde{r}} = \langle S, A, p, \widetilde{r}\rangle$,

$$\left|\rho_\pi(\mathcal{M}) - \rho_\pi(\mathcal{M}_{\widetilde{r}})\right| \;<\; \frac{\varepsilon}{c_r} \tag{11}$$

for each policy π. In the next step we want to modify the transition probabilities in $\mathcal{M}_{\widetilde{r}}$ so that for s, s' in the same block and for all blocks $S_i \in \widehat{S}$,

$$\sum_{s''\in S_i} \widetilde{p}_a(s, s'') = \sum_{s''\in S_i} \widetilde{p}_a(s', s''). \tag{12}$$

In order to attain this, we set for all $s \in S_j$, $s' \in S_i$, and all $a \in A$,

$$\widetilde{p}_a(s, s') \;:=\; p_a(s, s') + \frac{1}{|S_i|}\left(\frac{1}{|S_j|}\sum_{\bar{s}\in S_j}\sum_{s''\in S_i} p_a(\bar{s}, s'') - \sum_{s''\in S_i} p_a(s, s'')\right)$$

[2] Unfortunately, I haven't been aware of this reference until after submitting the final version of this paper to the proceedings of ALT 2007. Thus, this important reference is missing in the published paper.

(note that the $\widetilde{p}_a(s,\cdot)$ are indeed probability distributions for all $s \in S$), so that for s in any block S_j,

$$\begin{aligned}\sum_{s' \in S_i} \widetilde{p}_a(s,s') &= \sum_{s' \in S_i} p_a(s,s') + \frac{1}{|S_j|} \sum_{\bar{s} \in S_j} \sum_{s'' \in S_i} p_a(\bar{s},s'') - \sum_{s'' \in S_i} p_a(s,s'') \\ &= \frac{1}{|S_j|} \sum_{\bar{s} \in S_j} \sum_{s'' \in S_i} p_a(\bar{s},s''), \end{aligned} \tag{13}$$

independently of s, which entails (12). As $\widehat{S}$ is assumed to be a consistent ε-aggregation with respect to an adequate metric, we have by definition of $\widetilde{p}$ for transition probabilities $p(\cdot,\cdot)$, $\widetilde{p}(\cdot,\cdot)$ under any policy π and for $s \in S_j$, $s' \in S_i$,

$$\begin{aligned}|\widetilde{p}(s,s') - p(s,s')| \;&=\; \frac{1}{|S_i|} \cdot \left| \frac{1}{|S_j|} \sum_{\bar{s} \in S_j} \sum_{s'' \in S_i} p(\bar{s},s'') - \sum_{s'' \in S_i} p(s,s'') \right| \\ &\leq\; \frac{1}{|S_i|} \cdot \frac{1}{|S_j|} \sum_{\bar{s} \in S_j} \sum_{s'' \in S_i} \left| p(\bar{s},s'') - p(s,s'') \right| \;<\; \frac{\varepsilon}{c_p |S_i|}, \end{aligned}$$

so that for all $s \in S$,

$$\sum_{s' \in S} \left| \widetilde{p}(s,s') - p(s,s') \right| \;=\; \sum_{i=1}^{k} \sum_{s' \in S_i} \left| \widetilde{p}(s,s') - p(s,s') \right| \;<\; \sum_{i=1}^{k} \frac{\varepsilon}{c_p} \;=\; \frac{|\widehat{S}|}{c_p}\, \varepsilon.$$

Thus, by Theorem 3 we have for the difference of the average rewards of $\mathcal{M}_{\widetilde{r}}$ and $\widetilde{\mathcal{M}} := \langle S, A, \widetilde{p}, \widetilde{r} \rangle$ under some policy π (with respective stationary distributions μ_π and $\widetilde{\mu}_\pi$),

$$\begin{aligned}\left| \rho_\pi(\mathcal{M}_{\widetilde{r}}) - \rho_\pi(\widetilde{\mathcal{M}}) \right| \;&=\; \left| \sum_{s \in S} \big(\mu_\pi(s) - \widetilde{\mu}_\pi(s) \big)\, \widetilde{r}(s) \right| \;\leq\; \sum_{s \in S} \left| \mu_\pi(s) - \widetilde{\mu}_\pi(s) \right| \widetilde{r}(s) \\ &\leq\; \sum_{s \in S} \left| \mu_\pi(s) - \widetilde{\mu}_\pi(s) \right| \;=\; \left\| \mu_\pi - \widetilde{\mu}_\pi \right\|_1 \;<\; (\kappa_{\mathcal{M}_\pi} - 1) \frac{|\widehat{S}|}{c_p} \varepsilon. \end{aligned} \tag{14}$$

Now $\widetilde{\mathcal{M}}$ is a generalized block MDP with block set $\widehat{S}$, and by (10) and (13), its respective aggregated MDP is precisely $\widehat{\mathcal{M}}$. Analogously to the proof of Theorem 2, it follows from Lemma 2 that $\rho_\pi(\widehat{\mathcal{M}}) = \rho_{\pi^e}(\widetilde{\mathcal{M}})$ for all policies π on $\widehat{\mathcal{M}}$. Thus (11) and (14) yield

$$\begin{aligned}&\left| \rho_{\pi^e}(\mathcal{M}) - \rho_\pi(\widehat{\mathcal{M}}) \right| \;=\; \left| \rho_{\pi^e}(\mathcal{M}) - \rho_{\pi^e}(\widetilde{\mathcal{M}}) \right| \\ &\leq\; \left| \rho_{\pi^e}(\mathcal{M}) - \rho_{\pi^e}(\mathcal{M}_{\widetilde{r}}) \right| + \left| \rho_{\pi^e}(\mathcal{M}_{\widetilde{r}}) - \rho_{\pi^e}(\widetilde{\mathcal{M}}) \right| \;<\; \frac{\varepsilon}{c_r} + (\kappa_{\mathcal{M}_\pi} - 1) \frac{|\widehat{S}|}{c_p} \varepsilon. \quad \square \end{aligned}$$

Corollary 2. *Let π^* be an optimal policy on an MDP $\mathcal{M}$ with optimal average reward ρ^*, and let $\widehat{\pi}^*$ be an optimal policy with optimal average reward $\widehat{\rho}^*$ on a*

consistent ε-aggregation $\widehat{\mathcal{M}}$ of $\mathcal{M}$ with respect to an adequate metric. Then for $\kappa_{\mathcal{M}} := \max_\pi \kappa_{\mathcal{M}_\pi}$,

$$\textit{(i)} \quad |\rho^* - \widehat{\rho}^*| \;\leq\; \left(\frac{1}{c_r} + \frac{(\kappa_{\mathcal{M}} - 1)|\widehat{S}|}{c_p}\right)\varepsilon,$$

$$\textit{(ii)} \quad \rho^* \;\leq\; \rho_{\widehat{\pi}^{*e}}(\mathcal{M}) + \left(\frac{2}{c_r} + \frac{2(\kappa_{\mathcal{M}} - 1)|\widehat{S}|}{c_p}\right)\varepsilon.$$

Proof. Analogously to the proof of Corollary 1. □

Corollary 2 can be seen as a generalization of Lemma 4 of [2] to average reward MDPs. A similar result for average reward MDPs can be found in [9], see also footnote 2.

5 Dependence on the Mixing Time

5.1 Why Bounds are Worse in the Average Reward Case

The bounds obtained for ε-aggregation in the discounted case [3, 2] are basically of the form $\frac{\varepsilon}{1-\gamma} V_{\max}$, where $V_{\max}$ is the maximal possible discounted reward. Thus, on average one loses $\varepsilon V_{\max}$ reward in each step. This leads to the question whether the mixing time parameter in the obtained bounds for average reward MDPs is really necessary. It turns out that aggregation may go terribly wrong if mixing times are large.

Theorem 6. *For each $\varepsilon > 0$ and each $\delta \in (0, \varepsilon/2)$ there is an MDP $\mathcal{M}$ and an ε-aggregation $\widehat{\mathcal{M}}$ of $\mathcal{M}$ with respect to d_v, such that for some policy π on $\widehat{\mathcal{M}}$,*

$$|\rho_{\pi^e}(\mathcal{M}) - \rho_\pi(\widehat{\mathcal{M}})| \;\geq\; 1 - \delta.$$

Proof. Fix some $\varepsilon > 0$ and consider for $\delta \in (0, \varepsilon/2)$ the Markov chain $\mathcal{C}$ with $S = \{s_1, s_2, s_3\}$ and the following nonzero transition probabilities $p_{ij} := p(s_i, s_j)$,

$$p_{12} = 1 - \delta, \quad p_{13} = \delta, \quad p_{21} = p_{31} = \delta/n, \quad p_{22} = p_{33} = 1 - \delta/n,$$

where $n \in \mathbb{N}$. Then it is easy to check that we may ε-aggregate states s_1 and s_2 with respect to d_v so that we obtain a Markov chain $\widehat{\mathcal{C}}$ with states $S_1 = \{s_1, s_2\}$, $S_2 = \{s_3\}$ and transition probabilities

$$\widehat{p}(S_1, S_2) = \delta/2, \quad \widehat{p}(S_2, S_1) = \delta/n, \quad \widehat{p}(S_1, S_1) = 1 - \delta/2, \quad \widehat{p}(S_2, S_2) = 1 - \delta/n.$$

The original chain $\mathcal{C}$ has stationary distribution $\mu = \left(\frac{\delta}{n+\delta}, \frac{n-\delta n}{n+\delta}, \frac{\delta n}{n+\delta}\right)$, while the stationary distribution of $\widehat{\mathcal{C}}$ is $\widehat{\mu} = \left(\frac{2}{n+2}, \frac{n}{n+2}\right)$. Thus, for $n \to \infty$ one has $\widehat{\mu} \to (0, 1)$, while $\mu \to (0, 1-\delta, \delta)$. Thus any MDP whose induced Markov chain under some policy π is $\mathcal{C}$ satisfies the claim of the theorem, provided that π gives reward 1 in s_3 and reward 0 in s_1, s_2 (which is in accordance with ε-aggregation in respect to d_v). □

Thus the results for discounted MDPs are not transferable to the average reward case. Indeed, as shown in [10], the average reward ρ_π may be expressed via the discounted rewards $\rho_\pi^\gamma(s)$ as $\rho_\pi = (1-\gamma)\sum_s \mu_\pi(s)\,\rho_\pi^\gamma(s)$. This means that the stationary distribution μ_π under π plays an important role. The loss by aggregation remains small (just as in the discounted case) as long as $\widehat{\mu}$ approximates μ well, that is, $\widehat{\mu}(S_i) \approx \sum_{s\in S_i} \mu(s)$. The quality of approximation however can be estimated using the mixing time as Theorem 3 shows. Note that the mixing time in the example of Theorem 6 becomes arbitrarily large.

5.2 Alternative Perturbation Bounds

The perturbation bound for stationary distributions of Markov chains of Theorem 3, which we used in the proofs of Theorems 2 and 5, may be replaced with an arbitrary alternative perturbation bound of the form

$$\|\mu - \widetilde{\mu}\|_q \;\le\; \lambda\,\|P - \widetilde{P}\|_\infty.$$

There are several such bounds in the literature (for an overview see [11]). These differ from each other in at most two aspects, namely (i) the used norm q (which is either 1 or ∞) and (ii) the *conditioning number* λ. Obviously, bounds which hold for the ∞-norm instead of the 1-norm are impractical, as they would amount to an additional factor $|S|$ in the bounds of Theorems 2 and 5. Among the 1-norm bounds the conditioning number in terms of the mixing time used by Hunter has the advantage of being rather intuitive. However, there is little general knowledge about the size of the mixing time (cf. [5] for results in some special cases and also some comparison to other 1-norm conditioning numbers, which complements the overview given in [11]). Moreover, Seneta's *ergodicity coefficient* [12], which among the 1-norm conditioning numbers considered in [11] is the smallest, is in general also not larger than Hunter's mixing time parameter (see [13]), so that one may want to replace Theorem 3 with Seneta's perturbation bound [12]. Of course, this basically gives the same aggregation bounds, only that Hunter's mixing time parameter is replaced with Seneta's ergodicity coefficient.

6 Online Aggregation and Other Open Problems

Online Aggregation. Consider an agent who starts in an MDP unknown to her and tries to aggregate states while still collecting information about the MDP. Obviously, if she is given access to an adequate distance function, the aggregation may be done online. For given $\varepsilon > 0$ the most straightforward way to do this is to assign each newly visited state s to an existing block S_i if possible (i.e., if all states s' in S_i have distance $< \varepsilon$ to s), or otherwise create a new block $S_j \ni s$. This is an obvious sequential clustering algorithm (called e.g. BSAS in [14]). Also, Ferns et al. [3] suggest a similar approach for offline aggregation.

Unfortunately, even if the existence of a consistent ε-partition is guaranteed (which, as we have seen, need not be the case), in general this online aggregation

algorithm will give inconsistent ε-partitions. It is an interesting question whether there are more prospective algorithms for online aggregation.

More generally, a related open question is whether any online regret bounds are achievable for a combination of a suitable online aggregation algorithm with an online reinforcement learning algorithm (such as e.g. UCRL [15]). As it is of course hard to choose an appropriate ε in advance without having any information about the MDP at hand, one would need a mechanism which adapts the aggregation parameter ε to the MDP.

It may be relevant that generally, *optimal* aggregation is hard even if the MDP is known (cf. [2]). Although Even-Dar and Mansour consider discounted MDPs, their results hold generally, as the question is to find for given $\varepsilon > 0$ a minimal ε-adequate aggregation (see the modification of Definition 10).

Similarity of Actions. We have concentrated on MDPs with large state spaces. It is an interesting question whether an analogous approach will work for a similarity metric on actions, and in particular how the two approaches may be combined.

Relaxing Similarity. In many real-world problems one would want to relax the given similarity conditions. In particular, the idea that similar states shall lead to similar states under equivalent actions may not mean that states s, s' with $d(s, s') < \varepsilon$ will lead to states whose distance is $< \varepsilon$ as well. Rather one may e.g. demand that for some constant $c > 1$ the distance will be $< c\varepsilon$. Of course, under this generalized assumption no aggregation in the sense of a strict partition of the state space is possible anymore. Thus in order to deal with this setting, new methods will have to be developed.

Acknowledgements. The author would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported in part by the the Austrian Science Fund FWF (S9104-N04 SP4) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. We also acknowledge support by the PASCAL pump priming projects "Sequential Forecasting" and "Online Performance of Reinforcement Learning with Internal Reward Functions". This publication only reflects the authors' views.

References

1. Givan, R., Dean, T., Greig, M.: Equivalence notions and model minimization in Markov decision processes. Artif. Intell. **147**(1-2) (2003) 163–223
2. Even-Dar, E., Mansour, Y.: Approximate equivalence of Markov decision processes. In: Proc. 16th COLT, Springer (2003) 581–594
3. Ferns, N., Panangaden, P., Precup, D.: Metrics for finite Markov decision processes. In: Proc. 20th UAI, AUAI Press (2004) 162–169

4. Kemeny, J., Snell, J., Knapp, A.: Denumerable Markov Chains. Springer (1976)
5. Hunter, J.J.: Mixing times with applications to perturbed Markov chains. Linear Algebra Appl. **417** (2006) 108–123
6. Ortner, R.: Linear dependence of stationary distributions in ergodic Markov decision processes. Oper. Res. Lett. (2007) in press, doi:10.1016/j.orl.2006.12.001.
7. Puterman, M.L.: Markov Decision Processes. Discrete Stochastic Programming. Wiley (1994)
8. Leitgeb, H.: A new analysis of quasianalysis. J. Philos. Logic **36**(2) (2007) 181–226
9. Ren, K., Krogh, B.: State aggregation in Markov decision processes. In: Proceedings of the 41st IEEE Conference on Decision and Control, Volume 4, IEEE (2002) 3819–3824, doi:10.1109/CDC.2002.1184960
10. Singh, S.P., Jaakkola, T., Jordan, M.I.: Learning without state-estimation in partially observable Markovian decision processes. In: Proc. 11th ICML, Morgan Kaufmann (1994) 284–292
11. Cho, G.E., Meyer, C.D.: Comparison of perturbation bounds for the stationary distribution of a Markov chain. Linear Algebra Appl. **335** (2001) 137–150
12. Seneta, E.: Sensitivity analysis, ergodicity coefficients, and rank-one updates for finite Markov chains. In: Numerical solution of Markov chains. Dekker (1991) 121–129
13. Seneta, E.: Markov and the creation of Markov chains. In: MAM 2006: Markov Anniversary Meeting, Boson Books (2006) 1–20
14. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Third edn. Academic Press (2006)
15. Auer, P., Ortner, R.: Logarithmic online regret bounds for reinforcement learning. In: Proc. 19th NIPS, MIT Press (2006) 49–56