Value-Based Planning for Teams of Agents in Stochastic Partially Observable Environments
Oliehoek, F.A.

Citation for published version (APA):
Chapter 6

Lossless Clustering of Histories

The Generalized MAA* (GMAA*) algorithm introduced in Chapter 4 can find (optimal) solutions for Dec-POMDPs by repeatedly solving Bayesian games (BGs) for different stages. However, the cost of solving these BGs grows exponentially with the number of agents $n$ and doubly-exponentially with the horizon $h$. That is, for a BG for the last stage, the number of joint policies, and thus the cost of optimally solving it, is expressed by (5.0.1), repeated here for convenience:

$$O\left(|A_*|^n(|O_*|^{h-1})\right),$$

where $A_*$ and $O_*$ denote the largest individual action and observation sets. Chapter 5 introduced techniques to provide scaling with respect to the number of agents. This chapter addresses scaling with respect to the horizon.

In a BG for a stage $t$, an individual type corresponds to an action-observation history (AOH). However, given that a deterministic past joint policy $\varphi^t$ is followed, it also corresponds to a single observation history (OH), because the actions can be inferred from $\varphi^t$. The problem is that the number of individual types (i.e., the number of individual observation histories) grows exponentially over time and thus the BGs also grow exponentially.

To counter the exponential growth of the BGs this chapter proposes to cluster the individual AOHs in a way that does not compromise solution quality. It is straightforward to exploit this result in optimal policy search. In particular, Section 6.4 shows how to cluster histories within the GMAA* algorithm.

This chapter empirically demonstrates that the proposed technique can provide a speed-up of multiple orders of magnitude, allowing the optimal solution of significantly longer horizons. For instance, we solve the well-known benchmark decentralized tiger (DEC-Tiger) problem for horizon $h = 5$ (in which case there are $3.82c29$ joint policies) and the BROADCASTCHANNEL problem for horizons as large as $h = 20$. To the best of our knowledge, such results were not obtainable previously. Empirical analysis of the generality of the clustering method suggests that it may also be useful in other (approximate) Dec-POMDP solution methods.
6.1 Clustering Types in BGs

The idea of clustering histories in BGs is not new. Emery-Montemerlo et al. (2004) already proposed to prune types with low probabilities. In subsequent work Emery-Montemerlo et al. (2005) replaced this pruning by clustering types, based on the profiles of the payoff functions of the BGs, thereby increasing the quality of the found policies. However, since the payoff functions are heuristics, this method is somewhat ad-hoc: even when providing a bound on the error of clustering two types in a BG, as long as this bound is with respect to the heuristic payoff function, this will guarantee nothing with respect to the optimal solution of a Dec-POMDP. As such performing clustering based on the heuristic payoff function causes an increased dependence on heuristic, while providing no guarantees whatsoever.

This chapter also considers clustering of AOHs. In contrast, however, it does not consider a lossy clustering scheme based on the heuristic payoff function $\hat{Q}$ of the BGs. Rather, a criterion for clustering AOHs is introduced based on the probability these histories induce over histories of the other agents and over states. The nice thing of this criterion, which we refer to as probabilistic equivalence (PE), is that clustering histories that satisfy this criterion is lossless: the solution for the clustered BG can be used to construct the solution for the original BG and the values of the two BGs are identical. Thus, the criterion allows for clustering of AOHs in BGs that represent Dec-POMDPs without compromising solution quality, i.e., optimality is preserved.

The main contribution of this chapter is that it established that when two histories in a Dec-POMDP are PE, they can be clustered together without loss in value. In the proof, some other contributions are made. In the following we outline the proof and these other contributions.

Section 6.2 introduces two concepts: reduction of BGs through commitment and best-response equivalence for BGs. The former states that if an agent is committed to select the same action for two of its types, the BG can be reduced by clustering these types. The latter says when a rational agent is committed to select the same action for two of its types, namely when those types are guaranteed to have the same best-response action, i.e., if if the two types are best-response equivalent (BRE). These results, although applied in the context of Dec-POMDPs in this chapter, may be useful more generally.

Next, Section 6.3 applies these results in the Dec-POMDP context. In particular, it formally introduces probabilistic equivalence and demonstrates that if it holds for two histories, then they are BRE. This is proven by showing that if PE holds, the two conditions necessary for BRE, as will be identified by Lemma 6.1, hold.

6.2 Best-Response Equivalence for BGs

This section considers Bayesian games as introduced in Section 2.1.1.2 and investigates when it is possible to cluster individual types in BGs.
Theorem 6.1 (Reduction through commitment). Given that in a Bayesian game $B$ agent $i$ is committed to select a policy that assigns the same action for two of its types $\theta^a_i, \theta^b_i$, i.e., to select a policy $\beta_i$ such that

$$\beta_i(\theta^a_i) = \beta_i(\theta^b_i), \quad (6.2.1)$$

then the BG can be reduced to a smaller one without loss in value for any of the agents. I.e., the two types can be substituted by a new type $\theta^c_i$ such that

$$\forall \theta \neq \theta_i \quad \Pr(\theta^c_i, \theta \neq \theta_i) = \Pr(\theta^a_i, \theta \neq \theta_i) + \Pr(\theta^b_i, \theta \neq \theta_i) \quad (6.2.2)$$

$$\forall_j \forall a \quad u(\langle \theta^c_i, \theta \neq \theta_i \rangle, a) = \frac{\Pr(\theta^a_i, \theta \neq \theta_i) u(\langle \theta^a_i, \theta \neq \theta_i \rangle, a) + \Pr(\theta^b_i, \theta \neq \theta_i) u(\langle \theta^b_i, \theta \neq \theta_i \rangle, a)}{\Pr(\theta^a_i, \theta \neq \theta_i) + \Pr(\theta^b_i, \theta \neq \theta_i)} \quad (6.2.3)$$

The result is a new BG $B'$ in which the expected value is the same as in the original BG: $V^B = V^B$.

Proof. We show that the expected value of any joint policy $(\beta_i, \beta_{\neq i})$ that satisfies condition (6.2.1) is the same in both $B$ and $B'$. Using short-hand $a = \langle \beta_i(\theta_i), \beta_{\neq i}(\theta_{\neq i}) \rangle$,

$$V^B(\beta_i, \beta_{\neq i}) = \sum_{\theta \neq \theta_i} \left[ \Pr(\theta^a_i, \theta \neq \theta_i) u(\langle \theta^a_i, \theta \neq \theta_i \rangle, a) + \Pr(\theta^b_i, \theta \neq \theta_i) u(\langle \theta^b_i, \theta \neq \theta_i \rangle, a) \right]$$

$$+ \sum_{\theta_i \in \Theta_i \setminus \{\theta^a_i, \theta^b_i\}} \Pr(\theta_i, \theta \neq \theta_i) u(\langle \theta_i, \theta \neq \theta_i \rangle, a)$$

$$= \sum_{\theta \neq \theta_i} \left[ \Pr(\theta^c_i, \theta \neq \theta_i) u(\langle \theta^c_i, \theta \neq \theta_i \rangle, a) + \sum_{\theta_i \in \Theta_i \setminus \{\theta^c_i\}} \Pr(\theta_i, \theta \neq \theta_i) u(\langle \theta_i, \theta \neq \theta_i \rangle, a) \right]$$

$$= V^{B'}(\beta_i, \beta_{\neq i})$$

which is the expected value of $(\beta_i, \beta_{\neq i})$ as computed in the reduced BG. \qed

This theorem tells us that given that agent $i$ is committed to taking the same action for its types $\theta^a_i, \theta^b_i$, we can reduce the Bayesian game $B$ to a smaller one $B'$ and translate the joint BG-policy $\beta'$ found for $B'$ back to a joint BG-policy $\beta$ in $B$. This does not necessarily mean that $\beta = (\beta_i, \beta_{\neq i})$ also is a solution (Bayesian Nash-equilibrium) for $B$, because the best-response of agent $i$ against $\beta_{\neq i}$ may not select the same action for $\theta^a_i, \theta^b_i$. Rather $\beta_i$ is the best-response against $\beta_{\neq i}$ given that the same action needs to be taken for $\theta^a_i, \theta^b_i$. For instance, when $\theta^a_i, \theta^b_i$ are BRE as we detail below.

The preceding showed that reduction through clustering is possible if an agent is committed to select the same action for two of its types. In the following we will
identify when an agent is committed to select the same action for two of its types through the notion of best-response equivalence. I.e., the following demonstrates when the best-response for two types is the same. In a general BG, a best-response \( \beta^*_i \) for agent \( i \)'s type \( \theta_i \) against some fixed policy profile \( \beta_{\neq i} \) is given by

\[
\beta^*_i = \arg \max_{a_i} \sum_{\theta_i} \Pr(\theta_i) \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i}|\theta_i) u_i((\theta_i, \theta_{\neq i}), (\beta_i(\theta_i), \beta_{\neq i}(\theta_{\neq i}))),
\]

or, alternatively, we can compose \( \beta^*_i \) as the best response \( \beta^*_i(\theta_i) \) for each type \( \theta_i \): \( \beta^*_i(\theta_i) = \arg \max_{a_i} \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i}|\theta_i) u_i((\theta_i, \theta_{\neq i}), (a_i, \beta_{\neq i}(\theta_{\neq i}))). \)

It is this latter formulation we use in the following lemma, identifying best-response equivalence.

**Lemma 6.1 (Best-response equivalence).** When for two types \( \theta_{i,a}, \theta_{i,b} \) it holds that

\[
\forall \theta_{\neq i} \quad \Pr(\theta_{\neq i}|\theta_{i,a}) = \Pr(\theta_{\neq i}|\theta_{i,b}) \tag{6.2.4}
\]

and

\[
\forall a \forall \theta_{\neq i} \quad u(\theta_{i,a}, \theta_{\neq i}, a) = u(\theta_{i,b}, \theta_{\neq i}, a), \tag{6.2.5}
\]

then the best-response policy for agent \( i \) will always select the same action for \( \theta_{i,a}, \theta_{i,b} \).

**Proof.** We can simply derive

\[
\beta^*_i(\theta_{i,a}) = \arg \max_{a_i} \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i}|\theta_{i,a}) u(\theta_{i,a}, \theta_{\neq i}, a_i, \theta_{\neq i}) = \arg \max_{a_i} \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i}|\theta_{i,b}) u(\theta_{i,b}, \theta_{\neq i}, a_i, \theta_{\neq i}) \]

which is equal to \( \beta^*_i(\theta_{i,b}) \). \( \Box \)

**Remark.** This lemma states sufficient, but not necessary conditions for best response equivalence. This is easy to understand by considering a randomly generated BG with many types, but few actions per agent. Because the probabilities and utilities are randomly generated, conditions (6.2.4) and (6.2.5) typically will not hold. However, as there are many types and few actions, any policy (so also a best-response policy) will need to select the same action for many types.

By combination of Theorem 6.1 and Lemma 6.1, it is clear that individual types \( \theta_{i,a}, \theta_{i,b} \) can be clustered if the conditions in the lemma are satisfied. Note that although these results are presented in the context of identical payoff BGs, it is trivial to generalize them to BGs with individual payoff functions. As such, these results may perhaps also be used in more general methods for solving BGs.
6.3 Lossless Clustering in Dec-POMDPs

This section makes the bridge to the Dec-POMDP context. In particular it shows when two histories in a Dec-POMDP are BRE and can therefore be clustered together in a BG representing a stage of a Dec-POMDP.

6.3.1 Probabilistic Equivalence Criterion

A particular stage $t$ of a Dec-POMDP can be represented as a BG. For such a BG we can cluster two individual histories $\vec{\theta}_{t,i,a}, \vec{\theta}_{t,i,b}$ when they satisfy the probabilistic equivalence criterion as we define here.

**Criterion 6.1 (Probabilistic Equivalence).** Two AOHs $\vec{\theta}_{t,i,a}, \vec{\theta}_{t,i,b}$ for agent $i$ are probabilistically equivalent (PE) when the following holds:

\[
\forall \vec{\theta}_{\neq i}, \forall s \quad \Pr(s, \vec{\theta}_{\neq i} | \vec{\theta}_{t,i,a}) = \Pr(s, \vec{\theta}_{\neq i} | \vec{\theta}_{t,i,b}).
\]  

(6.3.1)

**Remark.** Alternatively, the criterion can be rewritten to the following two:

\[
\forall \vec{\theta}_{\neq i} \quad \Pr(\vec{\theta}_{\neq i} | \vec{\theta}_{t,i,a}) = \Pr(\vec{\theta}_{\neq i} | \vec{\theta}_{t,i,b}),
\]  

(6.3.2)

\[
\forall \vec{\theta}_{\neq i}, \forall s \quad \Pr(s | \vec{\theta}_{\neq i}, \vec{\theta}_{t,i,a}) = \Pr(s | \vec{\theta}_{\neq i}, \vec{\theta}_{t,i,b}).
\]  

(6.3.3)

These equations give a natural interpretation: the first says that the probability distribution over the other agents’ AOHs must be identical for both $\vec{\theta}_{t,i,a}$, $\vec{\theta}_{t,i,b}$. The second demands that the resulting joint beliefs are identical.

**Remark.** The above probabilities are not well defined without the initial state distribution $b^0$ and past joint policy $\varphi^t$. However, since we consider clustering of histories within a particular BG (for some stage $t$) and because this BG is constructed for a particular $b^0, \varphi^t$, they are implicitly specified. Therefore we drop these arguments, clarifying the notation.

**Remark.** Probabilities as defined in (6.3.1) appear somewhat similar to beliefs in POMDPs, but are substantially different. In a Dec-POMDP it is not possible for an agent to maintain beliefs as in POMDPs. The probabilities here are not sufficient statistics. Only a ‘multiagent belief’ specified over states and future policies of other agents has been shown to be a sufficient statistic (Hansen et al., 2004). Our notion of PE is specified over states and AOHs given only a past joint policy. Thus establishing conditions for equivalence in Dec-POMDPs is a non-trivial extension over the POMDP case.

Probabilistic equivalence has a convenient property: if it holds for a particular pair of histories, then it will also hold for all identical extensions of those histories, i.e., the property propagates forwards regardless of the policies the other agents use.

**Definition 6.1 (Identical extensions).** Given two AOHs $\vec{\theta}^t_{i,a}, \vec{\theta}^t_{i,b}$, their respective extensions $\vec{\theta}^{t+1}_{i,a} = (\vec{\theta}^t_{i,a}, a_i, o_i)$ and $\vec{\theta}^{t+1}_{i,b} = (\vec{\theta}^t_{i,b}, a'_i, o'_i)$ are called identical extensions if and only if $a_i = a'_i$ and $o_i = o'_i$. 
Lemma 6.2 (Propagation of PE). Given $\tilde{\theta}_{i,a}^t, \tilde{\theta}_{i,b}^t$ that are PE, regardless of $\beta_{\neq i}$ the policy the other agents use, identical extensions are also PE:

$$\forall a_i \forall o_{i+1} \forall \bar{a}_t \forall s_{t+1} \forall \bar{a}_{t+1}^s \Pr(s_{t+1}^+ | \bar{\theta}_{i,a}^t, a_i, o_{i+1}^t, \beta_{\neq i}^t) =$$

$$\Pr(s_{t+1}^+ | \bar{\theta}_{i,b}^t, a_i, o_{i+1}^t, \beta_{\neq i}^t)$$  (6.3.4)

Proof. Assume an arbitrary $a_i^t, o_{i+1}^t, \beta_{\neq i}^t, s_{t+1}$ and $\bar{\theta}_{i,a}^t = (\bar{\theta}_{i,a}^t, a_i^t, o_{i+1}^t)$. We have that

$$\Pr(s_{t+1}^+, \bar{\theta}_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t) =$$

$$\sum_{s^t} \Pr(o_{i+1}^t | a_i^t, \theta_{i,a}^t) \Pr(s_{t+1} | s^t, a_i^t, \theta_{i,a}^t, o_{i+1}^t, \beta_{\neq i}^t) \Pr(s_{t+1} | \theta_{i,a}^t, o_{i+1}^t)$$

Because we assumed an arbitrary $s_{t+1}^+, \bar{\theta}_{i,a}^t, o_{i+1}^t$, we have that

$$\forall s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t \Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t) =$$

$$\Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t)$$  (6.3.5)

In general we have that

$$\Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t) =$$

$$\frac{\Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t)}{\Pr(o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t)}$$

$$= \frac{\Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t)}{\sum_{s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t} \Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t)}$$

Now, because of (6.3.5), both the nominator and denominator are the same when substituting $\theta_{i,a}^t, \theta_{i,b}^t$ in this equation. This means we can conclude

$$\Pr(s_{t+1}^+, \theta_{i,a}^t, o_{i+1}^t | \theta_{i,a}^t, a_i^t, \beta_{\neq i}^t) = \Pr(s_{t+1}^+, \theta_{i,b}^t, o_{i+1}^t | \theta_{i,b}^t, a_i^t, \beta_{\neq i}^t)$$

Finally, because $a_i^t, o_{i+1}^t, \beta_{\neq i}^t, s_{t+1}$, $\theta_{i,a}^t$ were all arbitrarily chosen we can conclude (6.3.4).

6.3.2 Identical Values Allow Lossless Clustering of Histories

Since we want to show that two PE histories can be clustered under the optimal policy, we need to show (6.2.5) holds and thus that their optimal Q-values are the same.
Lemma 6.3 (Q\(^{\pi}\) equivalence). When two histories in a BG for a Dec-POMDP \(\vec{r}_{i,a}, \vec{r}_{i,b}\) satisfy Criterion 6.1, then they have equal Q-values according any joint policy \(\pi\)

\[
\forall \vec{\theta}_{\neq i} \forall a \quad Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}, a) = Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}, a).
\] (6.3.6)

Proof. The proof is by induction backwards in time (i.e., from the last time step \(t = h - 1\) to the first \(t = 0\)). However, to prove the induction step we employ Lemma 6.2, which ensures propagation forward through time of the PE criterion on identical extensions.

The base case is given by the last stage \(t = h - 1\) of the Dec-POMDP. In this case we have that

\[
\forall a \forall \vec{\theta}_{\neq i} \quad Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}, a) = \sum_{s \in S} R(s, a) \Pr(s | \vec{\theta}_{i,a}, \vec{\theta}_{\neq i}) =
\]

\[
\sum_{s \in S} R(s, a) \Pr(s | \vec{\theta}_{i,b}, \vec{\theta}_{\neq i}) = Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}, a)
\]

because of (6.3.3) in Criterion 6.1. For stages \(0 \leq t < h - 1\) the \(Q^{\pi}_{\pi}\) is given by

\[
Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,a}) = R(\vec{\theta}^{t}_{i,a}) + \sum_{o^{t+1} \in O} \Pr(o^{t+1} | \vec{\theta}^{t}_{i,a}) Q^{\pi}_{\pi}(\vec{\theta}^{t+1}_{i,a}, \pi(\vec{\theta}^{t+1}_{i,a})).
\]

The induction hypothesis is as follows: If at \(t + 1\) the criteria hold for any two \(\vec{\theta}^{t+1}_{i,a}, \vec{\theta}^{t+1}_{i,b}\), then they have equal Q-values:

\[
\forall \vec{\theta}^{t+1}_{i,a} \forall a^{t+1} \quad Q^{\pi}_{\pi}(\vec{\theta}^{t+1}_{i,a}, \vec{\theta}^{t+1}_{\neq i}, a^{t+1}) = Q^{\pi}_{\pi}(\vec{\theta}^{t+1}_{i,b}, \vec{\theta}^{t+1}_{\neq i}, a^{t+1}).
\] (6.3.7)

Assume: some stage \(0 \leq t < h - 1\), that the criteria hold for \(\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{i,b}\) and an arbitrary \(a = (a_{i}, a_{\neq i})\) and \(\vec{\theta}^{t}_{\neq i}\). Now we need to show that

\[
Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}, a) = Q^{\pi}_{\pi}(\vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}, a)
\] (6.3.8)

I.e.:

\[
R(\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}, a) + \sum_{o^{t+1} \in O} \Pr(o^{t+1} | \vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}, a) Q^{\pi}_{\pi}(\vec{\theta}^{t+1}_{i,a}, \pi(\vec{\theta}^{t+1}_{i,a})) =
\]

\[
R(\vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}, a) + \sum_{o^{t+1} \in O} \Pr(o^{t+1} | \vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}, a) Q^{\pi}_{\pi}(\vec{\theta}^{t+1}_{i,b}, \pi(\vec{\theta}^{t+1}_{i,b}))(6.3.9)
\]

where

\[
\vec{\theta}^{t+1}_{a} = ((\vec{\theta}^{t}_{i,a}, \vec{\theta}^{t}_{\neq i}), a, o^{t+1}) = (\vec{\theta}^{t}_{i,a}, a, o^{t+1}), (\vec{\theta}^{t}_{\neq i}, a, o^{t+1}) = (\vec{\theta}^{t+1}_{i,a}, \vec{\theta}^{t+1}_{\neq i})
\]

\[
\vec{\theta}^{t+1}_{b} = ((\vec{\theta}^{t}_{i,b}, \vec{\theta}^{t}_{\neq i}), a, o^{t+1}) = (\vec{\theta}^{t}_{i,b}, a, o^{t+1}), (\vec{\theta}^{t}_{\neq i}, a, o^{t+1}) = (\vec{\theta}^{t+1}_{i,b}, \vec{\theta}^{t+1}_{\neq i})
\]

To prove the equality of (6.3.9), we have to show that:
1. The immediate rewards are equal: 
\[ R(\tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b}, a) = R(\tilde{\theta}^t_{i,b}, \tilde{\theta}^t_{i}, a). \]
This clearly is the case (similar to the proof of the last stage).

2. \( \forall o^t+1 \) \( \Pr(o^t+1|\tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b}) = \Pr(o^t+1|\tilde{\theta}^t_{i,b}, \tilde{\theta}^t_{i}). \) Equal observation probabilities. This is also evident given that the criterion holds: if the underlying state distribution is the same, the next joint observation probabilities are also identical.

3. The relevant next-stage Q-values are identical. I.e.:
\[ \forall o^t+1 \forall a^t+1 Q^e_{\pi}(\tilde{\theta}^t_{i,a}, a^t+1) = Q^e_{\pi}(\tilde{\theta}^t_{i,b}, a^t+1). \tag{6.3.10} \]
To prove this, we show that the induction hypothesis applies: We can rewrite the demonstrandum (6.3.10) to
\[ \forall o^t+1 \forall o^t+1 \forall a^t+1 Q^e_{\pi}(\tilde{\theta}^t_{i,a}, a^t+1) = \]
\[ Q^e_{\pi}(\tilde{\theta}^t_{i,b}, a^t+1). \]
This is proven (by application of the induction hypothesis) if we can show that the criterion holds for \( \tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b} \). Since \( \tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b} \) are identical extensions of PE histories \( \tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b} \), they themselves are PE by application of Lemma 6.2. Therefore the induction hypothesis applies which means that (6.3.10) holds.

**Theorem 6.2** (Lossless clustering). When two histories \( \tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b} \) are PE, then they are best-response equivalent and can be clustered as one history without loss in value.

**Proof.** We first prove that PE implies BRE. The criterion itself entails (6.2.4). Lemma 6.3 asserts that, for a BG constructed using an optimal Q-value function \( Q^e_{\pi^*} \), (6.2.5) holds. Now, given that PE implies BRE, we can apply Theorem 6.1 to prove that \( \tilde{\theta}^t_{i,a}, \tilde{\theta}^t_{i,b} \) can be clustered without loss in value.

**Remark.** Again, the criterion gives a sufficient, but not necessary condition. In particular given a policy of the other agents, many types are BRE and can be clustered. However, as far as we know, only if the criterion holds we can guarantee that two histories have the same best-response against any policy of the other agents.

### 6.4 GMAA*-Cluster

Knowledge of which individual histories can be clustered together without loss of value may potentially be employed by many algorithms. In this paper, we focus on its application within the GMAA* framework.

Emery-Montemerlo et al. (2005) showed how clustering can be incorporated at every stage in their algorithm: when the BG for a stage \( t \) is constructed, first a
clustering of the individual histories (types) is performed and only afterward the (reduced) BG is solved. The same thing can be done within GMAA\(^*\), leading to an algorithm we dub GMAA\(^*\)-Cluster. In particular, GMAA\(^*\)-Cluster replaces the function ConstructAndSolveBG from Algorithm 4.1 with Algorithm 6.1.

The actual clustering is performed by Algorithm 6.2, which performs a pairwise comparison of all types of each agent to see if they satisfy the criterion. This means that

\[ O(|\Theta_i|^2) \]

are performed for each agent \(i\). If there is a large number of states some, efficiency may be gained by first checking (6.3.2) and then checking (6.3.3), rather than looping over all \(\langle s, \theta \neq i \rangle\) as is done in line 5.

Also note that the algorithm shown assumes that the heuristic used as the payoff function \(u\) is admissible (i.e., is an upper bound to the optimal value). Therefore, rather than using (6.2.3), we can take the lowest upper bound in line 15.\(^1\) In general this might increase the tightness of the heuristic, which can have a great effect on the performance as demonstrated in Chapter 4.

### 6.4.1 Bootstrapped Clustering

Because PE of AOHs propagates forwards (i.e., identical extensions of PE histories are also PE), we do not have to construct all \(|\mathcal{O}_i|^t\) possible AOHs at every stage \(t\) (given the past policy \(\varphi_i^t\) of agent \(i\)). Instead of clustering this exponentially growing set of types, we can simply extend the already clustered types of the previous stage’s BG, as shown in Algorithm 6.3.

That is, given \(\Theta_i\), the set of types of agent \(i\) at the previous stage \(t - 1\), and \(\beta_i^{t-1}\) the policy agent \(i\) took at that stage, the set of types at stage \(t\), \(\Theta_i^t\), can be constructed as

\[
\Theta_i^t = \{ \theta_i^t = (\theta_i, \beta_i^{t-1}(\theta_i), o_i^t) \mid \theta_i \in \Theta_i, o_i^t \in \mathcal{O}_i \}.
\]  
(6.4.1)

\(^1\)For the heuristics we employed there is no difference, because their heuristic value is also guaranteed to be the same if the criterion holds.
Algorithm 6.2 $BG = \text{ClusterBG}(BG)$

1: for each agent $i$ do 
2:    for each individual type $\theta_i \in BG, \Theta_i$ do 
3:        for each individual type $\theta'_i \in BG, \Theta_i$ do 
4:            isEquivalent ← true 
5:            for all $\langle s, \theta \neq_i \rangle$ do 
6:                if $\Pr(s, \theta \neq_i | \theta_i) \neq \Pr(s, \theta \neq_i | \theta'_i)$ then 
7:                    isEquivalent ← false 
8:                    break 
9:                end if 
10:            end for 
11:            if isEquivalent then 
12:                $BG, \Theta_i \leftarrow BG, \Theta_i \setminus \theta'_i$ \{Remove $\theta'_i$ from $BG$:\} 
13:                for each $a \in A$ do 
14:                    for all $\theta \neq_i$ do 
15:                        \{ take the lowest upper bound \} 
16:                        $u(\theta, \theta \neq_i, a) \leftarrow \min(u(\theta, \theta \neq_i, a), u(\theta'_i, \theta \neq_i, a))$ 
17:                        $\Pr(\theta, \theta \neq_i) \leftarrow \Pr(\theta, \theta \neq_i) + \Pr(\theta'_i, \theta \neq_i)$ 
18:                        $\Pr(\theta'_i, \theta \neq_i) \leftarrow 0$ 
19:                    end for 
20:                end if 
21:            end for 
22:        end for 
23:    end for 
24: end for 

This means that the size of this newly constructed set is

$$|\Theta'_i| = |\Theta_i| \cdot |O_i|$$ \hspace{1cm} (6.4.2)

If the typeset $\Theta_i$ at the previous stage $t - 1$ was much smaller than the set of all histories $|\Theta_i| \ll |O_i|^{t-1}$, then the new typeset $\Theta'_i$ is also much smaller: $|\Theta'_i| \ll |O_i|^t$. This way, we bootstrap the clustering at each stage and spend significantly less time clustering.

The above is possible only because we perform an exact, value preserving, clustering for which Lemma 6.2 tells us that identical extensions will also be clustered without loss in value. When performing the same procedure in a lossy clustering scheme (e.g., as in Emery-Montemerlo et al. 2005) errors might accumulate and thus it might be better to re-cluster from scratch at every stage. Still, this will mean that a resulting algorithm only has limited scalability. Since lossy clustering is beyond the scope of this chapter, only bootstrapped clustering is considered.

### 6.4.2 Complexity

Optimally solving a BG takes exponential time w.r.t. the number of types, as there are $O(|A|^n|\Theta|)$ joint BG-policies. Clustering involves a pairwise comparison of all
Algorithm 6.3 $BG^t = \text{ConstructExtendedBG}(BG^{t-1}, \beta^{t-1})$

1: $t \leftarrow BG^{t-1}.t + 1$
2: $pBG \leftarrow BG^{t-1}$
3: $pPol \leftarrow \beta^{t-1}$
4: for each agent $i$ do
5: $BG^t.\Theta_i = \text{ConstructExtendedTypeSet}(i)$
6: end for
7: for each joint type $\theta = (\theta^{t-1}, a^{t-1}, o^t) \in BG^t.\Theta$ do
8: for each state $s^t \in S$ do
9: Compute $\Pr(s^t|\theta)$
10: end for
11: $\Pr(\theta) \leftarrow \Pr(o^t|\theta^{t-1}, a^{t-1}) \Pr(\theta^{t-1})$
12: for each $a \in A$ do
13: $q \leftarrow \infty$
14: for each history $\tilde{\theta}^t$ represented by $\theta$ do
15: $q \leftarrow \min(q, \hat{Q}(\tilde{\theta}^t, a))$ \{ if $Q^* \leq \hat{Q}$ we can take the lowest upper bound \}
16: end for
17: $u(\theta, a) \leftarrow q$
18: end for
19: end for

6.5 Experiments

This section first presents a comparison of the optimal solution of several problems with and without clustering, followed by an analysis of the generality of lossless clustering, also for larger horizons for which optimal solutions are infeasible to compute.

GMAA*-Cluster is evaluated on a range of benchmark problems. All timing results mentioned in this chapter are CPU times with a resolution of 0.01s, and
Table 6.1: Results of GMAA* on several problems. Listed are the run times of regular GMAA* and GMAA*-Cluster, and the size of the BGs solved at each time step, with and without clustering.

<table>
<thead>
<tr>
<th>Dec-Tiger (QBG)</th>
<th>BroadcastChannel (QMDP)</th>
<th>GridSmall (QBG)</th>
<th>Cooperative Box Pushing (QMDP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>V*</td>
<td>T_{GMAA*} (s)</td>
<td>T_{cluster} (s)</td>
</tr>
<tr>
<td>2</td>
<td>4.0000</td>
<td>≤ 0.01</td>
<td>≤ 0.01</td>
</tr>
<tr>
<td>3</td>
<td>5.1908</td>
<td>0.02</td>
<td>≤ 0.01</td>
</tr>
<tr>
<td>4</td>
<td>4.8028</td>
<td>3.0694</td>
<td>1.50</td>
</tr>
<tr>
<td>5</td>
<td>7.0265</td>
<td>–</td>
<td>130.82</td>
</tr>
</tbody>
</table>

were obtained on 3.4GHz Intel Xeon processors with 2GB memory. The timings exclude time needed to parse the problem and compute the heuristic (which can be amortized).

### 6.5.1 Optimal Solutions using Clustering

For all considered problems we compared GMAA* against GMAA*-Cluster using the QBG or QMDP heuristic Oliehoek et al. (2008b), depending on problem size and planning horizon. Regardless of the particular heuristic, both methods compute an optimal policy, but we expect GMAA*-Cluster to be more efficient when lossless clustering is possible in the domain. The obtained results are shown in Table 6.1 and Table 6.2, which detail the optimal value V* and the running time T_{GMAA*} for
6.5 Experiments

Recycling Robots ($Q_{\text{MDP}}$)

| $h$ | $V^*$ | $T_{\text{GMAA}^*}(s)$ | $T_{\text{cluster}}(s)$ | $|BG^t|$ | $|cBG^t|$ |
|-----|-------|----------------------|----------------------|--------|--------|
| 2   | 6.8000 | $\leq 0.01$ | $\leq 0.01$ | 4   | 4.00   |
| 3   | 9.7647 | 0.02    | $\leq 0.01$ | 16  | 9.00   |
| 4   | 11.7264 | 23052.5 | 0.02    | 64  | 8.67   |
| 5   | 13.7643 | –      | 0.10    | 256 | 9.00   |
| 6   | 15.5760 | –      | 0.19    | 1024| 9.00   |
| 7   | 17.2126 | –      | 0.67    | 4096| 9.00   |
| 8   | 18.6839 | –      | 1.28    | 16384| 9.00 |
| 9   | 20.0085 | –      | 2.72    | 65536| 9.00 |
| 10  | 21.2006 | –      | 4.92    | 2.62e5| 9.00 |
| 11  | 22.2734 | –      | 9.83    | 1.05e6| 9.00 |
| 12  | 23.2390 | –      | 17.11   | 4.19e6| 9.00 |
| 13  | 24.1080 | –      | 30.61   | 1.68e7| 9.00 |
| 14  | 24.8901 | –      | 50.12   | 6.71e7| 9.00 |
| 15  | 25.5940 | –      | 81.46   | 2.68e8| 9.00 |

Hotel 1 ($Q_{BG}$)

| $h$ | $V^*$ | $T_{\text{GMAA}^*}(s)$ | $T_{\text{cluster}}(s)$ | $|BG^t|$ | $|cBG^t|$ |
|-----|-------|----------------------|----------------------|--------|--------|
| 2   | 9.5000 | $\leq 0.01$ | 0.02    | 16   | 4.00   |
| 3   | 15.7047 | –      | 0.07    | 256  | 16.00  |
| 4   | 20.1125 | –      | 1.37    | 4096 | 32.00  |

FireFighting ($n_h = 3, n_f = 3$) ($Q_{BG}$)

| $h$ | $V^*$ | $T_{\text{GMAA}^*}(s)$ | $T_{\text{cluster}}(s)$ | $|BG^t|$ | $|cBG^t|$ |
|-----|-------|----------------------|----------------------|--------|--------|
| 2   | -4.3825 | 0.03    | 0.03    | 4     | 4.00   |
| 3   | -5.7370 | 0.91    | 0.70    | 16    | 16.00  |
| 4   | -6.5789 | 5605.3  | 5823.5  | 64    | 64.00  |

Table 6.2: Results of GMAA* on several problems. Listed are the run times of regular GMAA* and GMAA*-Cluster, and the size of the BGs solved at each time step, with and without clustering.

GMAA* and $T_{\text{cluster}}$ for GMAA*-Cluster. Entries marked ‘−’ indicate that no solution was found within 8 hours. Furthermore, the tables list the number of joint types in the BGs constructed for the last stage without clustering, $|BG^t|$, and with, $|cBG^t|$. The former is constant while the latter is an average, as the algorithm can form BGs for different past policies, leading to clusterings of different sizes. For the Dec-Tiger problem, the solution time needed by GMAA*-Cluster is more than 3 orders of magnitude less for horizon $h = 4$. For $h = 5$ this test problem has 3.82e29 joint policies. To our knowledge, no other method has been able to optimally solve $h = 5$ Dec-Tiger. GMAA*-Cluster, however, is able to solve Dec-Tiger for $h = 5$ in reasonable time.

For the FireFighting problem, no lossless clustering is possible at any stage, and as such, we incur some overhead for the clustering. This is clearly shown for $h = 4$. For horizon 3, GMAA*-Cluster is actually a bit faster. Analysis revealed that for this horizon the cost of attempting to cluster is negligible. GMAA*-Cluster is faster because constructing the BGs using bootstrapping from the previous BG
takes less time than constructing a BG from scratch.

For GridSmall, Cooperative Box Pushing, and Hotel 1 the results are comparable to those for Dec-Tiger: substantial clustering is possible, resulting in significant speedups. Because the solution of BGs takes time exponential in their size, even small reductions in size would yield a big increase in efficiency. The substantial amounts of clustering found in these problems, therefore allow optimal solutions for longer horizons than have have been presented before.

For BroadcastChannel, GMAA*-Cluster achieves an even more dramatic increase in performance, allowing the solution of up to horizon \( h = 25 \). Analysis reveals that the BGs constructed for all stages are fully clustered: they contain only one type for each agent. Consequently, the time needed to solve each BG does not grow with the horizon. The solution time, however, still increases super-linear because of the increased amount of backtracking and memory management. The Recycling Robots problem can also be clustered to a relatively constant number of approximately 9 joint types per stage, allowing for optimal solving to high horizons. Both the BroadcastChannel and Recycling Robots problem run out of (2GB of) memory for higher horizons.

The fact that the BroadcastChannel problem exhibits full clustering can be explained as follows. When constructing a BG for \( t = 1 \), there is only one joint type for the previous BG, so given \( \beta^0 \), the solution for the previous BG, there is no uncertainty with respect to the previous joint action \( a^0 \). The crucial property of BroadcastChannel is that the (joint) observation tells us nothing about the new state, but only about what joint action was taken (e.g., ‘collision’ if both agent chose to ‘send’). As a result, the observation does not convey any information and the different individual histories can be clustered. In a BG constructed for stage \( t = 2 \), there will again be only one joint type in the previous game. Therefore, given the past policy, the actions of the other agents can be perfectly predicted. Again the observation will convey no information so this process repeats. Consequently, the problem can be considered a special form of a non-observable Dec-POMDP; lossless clustering automatically exploits this property.

Another special class of problems that exhibits full clustering are those with a known start state and deterministic actions. Again in this case, the observations convey no information (because we can perfectly predict everything), and all histories can be clustered. This can be described as a special case of a fully observable problem which clustering automatically exploits.

The FireFighting problem does not allow any clustering. This can be understood as follows: given a (heuristically) good past joint policy, each agent typically visits different houses (cf. Figure 2.4 on page 28). As such, each different observation history will typically induce a different belief over the global state.

The other problems are harder to analyze. In Dec-Tiger a key property is that opening the door resets the problem. Such resets invalidate the history, allowing for clustering. Another factor is that the observations are taken independently given the new state only. I.e., \( \Pr(o|a,s') = \Pr(o_1|s') \Pr(o_2|s') \), which means that all information regarding the history of the other agent is obtained through estimation of the state.
6.6 Conclusions

6.5.2 General Clustering Performance

The reduction in BG-size in GMAA*-Cluster leads to significant gains in efficiency, showing that heuristically high-ranked partial policies lead to BGs that allow for much clustering. To test the general applicability of the clustering method, we investigated how much clustering can be done in BGs constructed for random past policies. If substantial clustering is possible on random policies, not just those encountered by GMAA*-Cluster, then the approach may be useful for a much broader set of methods. The results are shown in Figure 6.1, which shows the median number of joint types $|cBG^t|$ in the Bayesian games (constructed for 1,000 random past policies) for different stages after clustering.

The FireFighting problem, which could not be clustered when searching for an optimal policy, does allow for some clustering given randomly selected policies (Figure 6.1g). In both the Recycling Robots and the Hotel 1 problem the growth in BG size appears to stabilize, while in Dec-Tiger, GridSmall, and Cooperative Box Pushing $|cBG^t|$ keeps growing in the planning horizon. Even so, $|BG^t|$ grows faster resulting in high clustering ratios also for these problems.

These experiments imply that the proposed clustering technique can provide significantly smaller policy representations without loss of value at a relatively low computational cost, for the benefit of optimal and approximate algorithms alike. Also this technique gives insight into how many future policies an agent should consider: if at some stage and given a past policy an agent has only $k$ types, this means that it maximally needs to consider $k$ future policies from that situation. The memory bounded dynamic programming (MBDP) algorithm and its variants (e.g., Seuken and Zilberstein, 2007b; Carlin and Zilberstein, 2008)), discussed in Section 2.6.5.1, have a parameter controlling the number of future policies considered, but until now there has been no principled way of estimating good values for this parameter. As such we expect that this clustering technique can have a big impact on new and existing, exact and approximate algorithms.

6.6 Conclusions

This chapter introduced a method for lossless clustering of action-observation histories in Dec-POMDPs, which can be applied in GMAA* policy search for Dec-POMDPs via Bayesian games. Rather than applying an ad-hoc clustering of these BGs, a probabilistic equivalence criterion was identified that guarantees that, given a particular past joint policy $\varphi^t$, two action-observation histories $\theta^t_i$ of agent $i$ at stage $t$ have the same optimal Q-values and therefore can be clustered without loss in solution quality. Empirical evaluation of GMAA* demonstrated that for several domains speedups of multiple orders of magnitude are achieved by clustering. We also investigated the amount of clustering possible for random past policies $\varphi^t$, the result of which suggests that our clustering methods may also be exploited in other algorithms, such as IMBDP (Seuken and Zilberstein, 2007b).
Figure 6.1: Empirical clustering performance given random joint policies, for several problems, based on 1,000 independent samples. Plots (a)–(f) show the median size of the Bayesian games at each stage after clustering $|BG^t|$, and the errorbars show the 0.25 and 0.75-quantile. Table (g) shows their median clustering ratio $\frac{|BG^t|}{|cBG^t|}$ for the last time step tested.


6.7 Discussion and Future Work

The empirical results shown in this chapter demonstrate that lossless clustering offers dramatic performance gains on a diverse set of problems. However, since some domains cannot be clustered in this way, it remains unclear in exactly what types of problems lossless clustering is effective. This is a hard question, as it requires an analysis of the subclasses of Dec-POMDPs, a matter about which relatively few results are known. Most research has focused on analysis of methods, rather than of properties of Dec-POMDP problems, notable exceptions being (Pynadath and Tambe, 2002a; Goldman and Zilberstein, 2004). Although a detailed analysis is beyond the scope of this chapter, some observations based on empirical results have been provided.

Boulaertas and Chaib-draa (2008) also present an algorithm that improve efficiency of optimal solutions by a form of compression. The performance of their algorithm, however, stays behind when compared to GMAA*-Cluster. Although a more careful analysis is needed, there are two main reasons that can explain this. First, the compression of Boulaertas and Chaib-draa works on the exponentially larger space of policies, while GMAA*-Cluster works on an exponentially smaller space of histories. Second, GMAA*-Cluster can exploit knowledge of the initial state distribution $b^{0}$.

The criterion for clustering is quite strict and there will also be many problems in which little or no lossless clustering is possible. In the future, we plan to consider approximations for such cases. In particular, one idea is to cluster approximately PE histories, e.g., if Kullback-Leibler divergence is below some threshold. Another idea is to cluster histories that induce the same individual belief over states:

$$\Pr(s|\tilde{\theta}_i) = \sum_{\tilde{\theta} \neq i} \Pr(s, \tilde{\theta} \neq i | \tilde{\theta}_i). \quad (6.7.1)$$

Such individual beliefs literally summarize the criterion and may therefore perform quite well in practice. Further investigation is needed to determine for which classes of problems such approximations might work.