Universiteit van Amsterdam

IAS technical report IAS-UVA-07-03

Properties of the Q_{BG}-value function

Frans A. Oliehoek[1], **Nikos Vlassis**[2], **and Matthijs T.J. Spaan**[3]

[1]ISLA, University of Amsterdam, The Netherlands
[2]Dept. of Production Engineering and Management, Technical University of Crete, Greece
[3]Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal

In this technical report we treat some properties of the recently introduced Q_{BG}-value function. In particular we show that it is a piecewise linear and convex function over the space of joint beliefs. Furthermore, we show that there exists an optimal infinite-horizon Q_{BG}-value function, as the Q_{BG} backup operator is a contraction mapping. We conclude by noting that the optimal Dec-POMDP Q-value function cannot be defined over joint beliefs.

Keywords: Multiagent systems, Dec-POMDPs, planning, value functions.

IAS
intelligent autonomous systems

Contents

Intelligent Autonomous Systems
Informatics Institute, Faculty of Science
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands

Tel (fax): +31 20 525 7461 (7490)
http://www.science.uva.nl/research/ias/

Corresponding author:

F.A. Oliehoek
tel: +31 20 525 7524
faolieho@science.uva.nl
http://www.science.uva.nl/~faolieho/

1 Introduction

The decentralized partially observable Markov decision process (Dec-POMDP) [1] is a generic framework for multiagent planning in a partially observable environment. It considers settings where a team of agents have to cooperate as to maximize some performance measure, which describes the task. The agents, however, cannot fully observe the environment, i.e., there is state uncertainty: each agent receives its own observations which provide a clue regarding the true state of the environment.

Emery-Montemerlo et al. [3] proposed to use a series of Bayesian games (BG) [6] to find an approximate solution for Dec-POMDPs, by employing a heuristic payoff function for the BGs. In previous work [5], we extended this modeling to the exact setting by showing that there exist an optimal Q-value function Q^* that, when used as the payoff function for the BGs, yields the optimal policy. We also argued that computing Q^* is hard and introduce Q_{BG} as a new approximate Q-value function that is a tighter upper bound to Q^* than previous approximate Q-value functions. Apart from its use as an approximate Q-value function for (non-communicative) Dec-POMDPs [5], the Q_{BG}-value function can also be used in communicative Dec-POMDPs: when assuming the agents in a Dec-POMDP can communicate freely, but that this communication is delayed by one time step, the Q_{BG}-value function is optimal.

In this report we treat several properties of the Q_{BG}-value function. We show that for a finite horizon, the Q_{BG} Q-value function $Q_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a})$, corresponds with a value function over the joint belief space $Q_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a})$ and that it is *piecewise linear and convex (PWLC)*.

For the infinite-horizon case, we also show that we can define a Q_{BG} backup operator, and that the operator is a contraction mapping. As a result we can conclude the existence of an optimal $\mathrm{Q}^*_{\mathrm{BG}}$ for the infinite horizon.

First, we further formalize Dec-POMDPs and relevant notions, then section 3 treats the finite horizon: 3.1 shows that Q_{BG} is a function over the belief space and section 3.2 we prove that this function is PWLC. In section 4 we treat the infinite-horizon case: section 4.1 shows that in this case joint beliefs are also a sufficient statistic. Section 4.2 shows how the Q_{BG} functions can be altered to form a backup operator for the infinite-horizon case and that this operator is a contraction mapping. Finally, in Section 5 we prove that the optimal Dec-POMDP Q-value function cannot be defined over joint beliefs.

2 Model and definitions

As mentioned, we adopt the Dec-POMDP framework [1].

Definition 2.1 A *decentralized partially observable Markov decision process (Dec-POMDP)* with m agents is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, O \rangle$ where:

- $\mathcal{S}$ is a finite set of states.
- The set $\mathcal{A} = \times_i \mathcal{A}_i$ is the set of *joint actions*, where $\mathcal{A}_i$ is the set of actions available to agent i. Every time step one joint action $\mathbf{a} = \langle a_1, ..., a_m \rangle$ is taken.[1]
- T is the transition function, a mapping from states and joint actions to probability distributions over next states: $T : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$.[2]
- R is the reward function, a mapping from states and joint actions to real numbers: $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

[1] Unless stated otherwise, subscripts denote agent indices.

[2] We use $\mathcal{P}(X)$ to denote the infinite set of probability distributions over the finite set X.

- $\mathcal{O} = \times_i \mathcal{O}_i$ is the set of joint observations, with $\mathcal{O}_i$ the set of observations available to agent i. Every time step one joint observation $\mathbf{o} = \langle o_1, ..., o_m \rangle$ is received.

- O is the observation function, a mapping from joint actions and successor states to probability distributions over joint observations: $O : \mathcal{A} \times \mathcal{S} \to \mathcal{P}(\mathcal{O})$.

Additionally, we assume that $b^0 \in \mathcal{P}(\mathcal{S})$ is the initial state distribution at time $t = 0$.

The planning problem is to compute a plan, or *policy,* for each agent that is optimal for a particular number of time-steps h, also referred to as the *horizon* of the problem. A common optimality criterion is the expected cumulative (discounted) future reward:

$$E\left(\sum_{t=0}^{h-1} \gamma^t R(t)\right). \tag{2.1}$$

The horizon h can be assumed to be finite, in which case the discount factor γ is generally set to 1, or one can optimize over an infinite horizon, in which case $h = \infty$ and $0 < \gamma < 1$ to ensure that the above sum is bounded.

In a Dec-POMDP, policies are mappings from a particular history to actions. Here we introduce a very general form of history.

Definition 2.2 The *action-observation history for agent i*, $\vec{\theta}_i^t$, is the sequence of actions taken and observations received by agent i until time step t:

$$\vec{\theta}_i^t = \left(a_i^0, o_i^1, a_i^1, ..., a_i^{t-1}, o_i^t\right). \tag{2.2}$$

The *joint action-observation history* is a tuple with the action-observation history for all agents $\vec{\theta}^t = \left\langle \vec{\theta}_1^t, ..., \vec{\theta}_m^t \right\rangle$. The set of all action-observation histories for agent i at time t is denoted $\vec{\Theta}_i$.

In the $\mathrm{Q_{BG}}$ setting, at a time step t the previous joint action-observation history $\vec{\theta}^{t-1}$ is assumed common knowledge, as the one-step-delayed communication of $\mathbf{o}^{t-1}$ has arrived. When planning is performed off-line, the agents know each others policies and $\mathbf{a}^{t-1}$ can be deduced from $\vec{\theta}^{t-1}$. The remaining uncertainty is regarding the last joint observation $\mathbf{o}^t$. This situation can be modeled using a *Bayesian game (BG)* [6]. In this case the *type* of agent i corresponds to its last observation $\theta_i \equiv o_i^t$. $\Theta = \times_i \Theta_i$ is the set of joint types, here corresponding with the set of joint observations $\mathcal{O}$, over which a probability function $P(\Theta)$ is specified, in this case $P(\theta) \equiv P(\mathbf{o}^t | \vec{\theta}^{t-1}, \mathbf{a}^{t-1})$. Finally, the BG also specifies a payoff function $u(\theta, \mathbf{a})$ that maps joint types and actions to rewards.

A joint BG-policy is a tuple $\beta = \langle \beta_1, ..., \beta_m \rangle$, where the individual policies are mappings from types to actions: $\beta_i : \Theta_i \to \mathcal{A}_i$. The solution of a BG with identical payoffs for all agents is given by the optimal joint BG-policy β^*:

$$\beta^* = \arg\max_{\beta} \sum_{\theta \in \Theta} P(\theta) u(\theta, \beta(\theta)), \tag{2.3}$$

where $\beta(\theta) = \langle \beta_1(\theta_1), ..., \beta_m(\theta_m) \rangle$ is the joint action specified by β for joint type θ. In this case, for a particular joint action-observation history $\vec{\theta}^t$, the agents know $\vec{\theta}^{t-1}$ and $\mathbf{a}^{t-1}$ and they solve the corresponding BG:

$$\beta^*_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1} \rangle} = \arg\max_{\beta_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1} \rangle}} \sum_{\mathbf{o}^t} P(\mathbf{o}^t | b^{\vec{\theta}^{t-1}}, \mathbf{a}^{t-1}) u_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1} \rangle}(\mathbf{o}^t, \beta_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1} \rangle}(\mathbf{o}^t)), \tag{2.4}$$

When defining $u_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1} \rangle}(\mathbf{o}^t, \mathbf{a}^t) \equiv Q^*_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a}^t)$, the $\mathrm{Q_{BG}}$-value function is the optimal payoff function [5]. It is given by:

$$Q^*_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a}) = R(\vec{\theta}^t, \mathbf{a}) + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|\vec{\theta}^t, \mathbf{a}) Q^*_{\mathrm{B}}(\vec{\theta}^{t+1}, \beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle}(\mathbf{o}^{t+1})), \tag{2.5}$$

where $\beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle} = \left\langle \beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle,1}(o_1^{t+1}), ..., \beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle,m}(o_m^{t+1}) \right\rangle$ is a tuple of individual policies $\beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle,i} : \mathcal{O}_i \to \mathcal{A}_i$ for the BG played for $\vec{\theta}^t, \mathbf{a}$, and where $R(\vec{\theta}^t, \mathbf{a}) = \sum_s R(s, \mathbf{a}) P(s|\vec{\theta}^t)$ is the expected immediate reward.

Note that the $\mathrm{Q_{BG}}$-setting is quite different from the standard Dec-POMDP setting, as shown in [5]. In this latter case, rather than solving a BG for each $\vec{\theta}^{t-1}$ and $\mathbf{a}^{t-1}$ (i.e., (2.4)) the agents solve a BG for each time step $0, 1, ..., h-1$:

$$\beta^{t,*} = \arg\max_{\beta^t} \sum_{\vec{\theta}^t \in \vec{\Theta}^t_{\pi^*}} P(\vec{\theta}^t) Q^*(\vec{\theta}^t, \beta^t(\vec{\theta}^t)). \tag{2.6}$$

When the the summation is over $\vec{\Theta}^t_{\pi^*}$: all joint action-observation histories that are consistent with the optimal joint policy π^*[3], and when the BGs use the optimal Q-value function:

$$Q^*(\vec{\theta}^t, \mathbf{a}) = R(\vec{\theta}^t, \mathbf{a}) + \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|\vec{\theta}^t, \mathbf{a}) Q^*(\vec{\theta}^{t+1}, \pi^*(\vec{\theta}^{t+1})), \tag{2.7}$$

then, solving the BGs for time step $0, 1, ..., h-1$ will yield the optimal policy π^*, i.e., $\pi^{t,*} \equiv \beta^{t,*}$.

3 Finite horizon

In this section we will consider several properties of the finite-horizon $\mathrm{Q_{BG}}$-value function.

3.1 $\mathrm{Q_{BG}}$ is a function over the joint belief space

In a single agent POMDP, a *belief* b is a probability distribution over states that forms a sufficient statistic for the decision process. In a Dec-POMDP we use the term *joint belief* and write $b^{\vec{\theta}^t} \in \mathcal{P}(\mathcal{S})$ for the probability distribution over states induced by joint action-observation history $\vec{\theta}^t$. Here we show that the $\mathrm{Q_{BG}}$ value function $Q_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a})$ corresponds with a Q-value function over the space of joint beliefs $b^{\vec{\theta}^t}$.

Lemma 3.1 *The Q_{BG}-value function* (2.5) *is a function over the joint belief space, I.e., it is possible convert* (2.5) *to a Q-value function over this joint belief space by substituting the action-observation histories by their induced joint beliefs:*

$$Q^*_B(b^{\vec{\theta}^t}, \mathbf{a}) = R(b^{\vec{\theta}^t}, \mathbf{a}) + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|b^{\vec{\theta}^t}, \mathbf{a}) Q^*_B(b^{\vec{\theta}^{t+1}}, \beta_{\langle \vec{\theta}^t, \mathbf{a}\rangle}(\mathbf{o}^{t+1})), \tag{3.1}$$

where $b^{\vec{\theta}^t}$ denotes the joint belief induced by $\vec{\theta}^t$.

Proof First we need to show that there exists exactly one joint belief over states $b^{\vec{\theta}^t} \in \mathcal{P}(\mathcal{S})$ for each joint-action observation history $\vec{\theta}^t$. This is almost trivial: using Bayes' rule we can calculate the joint belief $b^{\vec{\theta}^{t+1}}$ resulting from $b^{\vec{\theta}^t}$ by $\mathbf{a}$ and $\mathbf{o}^{t+1}$ by:

$$\forall_{s^{t+1}} \qquad b^{\vec{\theta}^{t+1}}(s^{t+1}) = \frac{P(s^{t+1}, \mathbf{o}^{t+1}|b^{\vec{\theta}^t}, \mathbf{a})}{P(\mathbf{o}^{t+1}|b^{\vec{\theta}^t}, \mathbf{a})}. \tag{3.2}$$

Because we assume only one initial belief, there is exactly one joint belief $b^{\vec{\theta}^t}$ for each $\vec{\theta}^t$.

[3] I.e., the action-observation histories that specify the same actions for all observation histories as π^*.

Of course the converse is not necessarily true: a particular distribution over states can correspond to multiple joint action-observation histories. Therefore, to show that the conversion from $Q_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a})$ to $Q_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a})$ is possible we will need to show that it is impossible that two different joint action-observation histories $\vec{\theta}^{t,a}, \vec{\theta}^{t,b}$ corresponds to the same belief , but have different Q_{BG} values. I.e., we have to show that if $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$ then

$$\forall_{\mathbf{a}} \quad Q^*_{\mathrm{B}}(\vec{\theta}^{t,a}, \mathbf{a}) = Q^*_{\mathrm{B}}(\vec{\theta}^{t,b}, \mathbf{a}) \tag{3.3}$$

holds.

We give a proof by induction, the base case is given by the last time step $t = h - 1$. In this case (2.5) reduces to:

$$Q^*_{\mathrm{B}}(\vec{\theta}^t, \mathbf{a}) = R(\vec{\theta}^t, \mathbf{a}) = \sum_s R(s, \mathbf{a}) b^{\vec{\theta}^t}(s). \tag{3.4}$$

Clearly, if $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$ then (3.3) holds. Therefore the base case holds. Now we need to show that if $b^{\vec{\theta}^{t+1,a}} = b^{\vec{\theta}^{t+1,b}}$ implies $Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,a}, \mathbf{a}) = Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,b}, \mathbf{a})$, then it should also hold that $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$ implies $Q^*_{\mathrm{B}}(\vec{\theta}^{t,a}, \mathbf{a}) = Q^*_{\mathrm{B}}(\vec{\theta}^{t,b}, \mathbf{a})$.

In the base case, the immediate rewards $R(\vec{\theta}^{t,a}, \mathbf{a})$ and $R(\vec{\theta}^{t,b}, \mathbf{a})$ are equal when $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$. Therefore we only need to show that the future reward is also equal. I.e., we need to show that, if $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$, it holds that

$$\max_{\beta_{\langle \vec{\theta}^{t,a}, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | \vec{\theta}^{t,a}, \mathbf{a}) Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,a}, \beta_{\langle \vec{\theta}^{t,a}, \mathbf{a} \rangle}(\mathbf{o}^{t+1})) =$$
$$\max_{\beta_{\langle \vec{\theta}^{t,b}, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | \vec{\theta}^{t,b}, \mathbf{a}) Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,b}, \beta_{\langle \vec{\theta}^{t,b}, \mathbf{a} \rangle}(\mathbf{o}^{t+1})), \tag{3.5}$$

given that $b^{\vec{\theta}^{t+1,a}} = b^{\vec{\theta}^{t+1,b}}$ implies $Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,a}, \mathbf{a}) = Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,b}, \mathbf{a})$.

Because $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$, we know that for each $\mathbf{a}, \mathbf{o}^{t+1}$ the resulting beliefs will be the same $b^{\vec{\theta}^{t+1,a}} = b^{\vec{\theta}^{t+1,b}}$. The induction hypothesis says that the Q_{BG}-values of the resulting joint beliefs are also equal in that case, i.e., $\forall_{\mathbf{a}} \; Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,a}, \mathbf{a}) = Q^*_{\mathrm{B}}(\vec{\theta}^{t+1,b}, \mathbf{a})$. Also it is clear that the probabilities of joint observations are equal $\forall_{\mathbf{o}^{t+1}} \; P(\mathbf{o}^{t+1} | \vec{\theta}^{t,a}, \mathbf{a}) = P(\mathbf{o}^{t+1} | \vec{\theta}^{t,b}, \mathbf{a})$.

Therefore, the future rewards for $\vec{\theta}^{t,a}$ and $\vec{\theta}^{t,b}$ as shown by (3.5) must be equal: they are defined as the value of the optimal solution to identical Bayesian games (meaning BGs with the same probabilities and payoff function). □

3.2 Q_{BG} is PWLC over the joint belief space

Here we prove that the Q_{BG}-value function is PWLC. The proof is a variant of the proof that the value function for a POMDP is PWLC [7].

Theorem 3.1 *The Q_{BG}-value function for a finite horizon Dec-POMDP with 1 time step delayed, free and noiseless communication, as defined in* (3.1) *is piecewise-linear and convex (PWLC) over the joint belief space.*

Proof The proof is by induction. The base case is the last time step $t = h - 1$. For the last time step (3.1) reduces to:

$$Q^*_{\mathrm{B}}(b^{\vec{\theta}^{h-1}}, \mathbf{a}) = R(b^{\vec{\theta}^{h-1}}, \mathbf{a}) = \sum_s R(s, \mathbf{a}) b^{\vec{\theta}^{h-1}}(s) = R_{\mathbf{a}} \cdot b^{\vec{\theta}^{h-1}}, \tag{3.6}$$

where $R_{\mathbf{a}}$ is the immediate reward vector for joint action $\mathbf{a}$, directly given by the immediate reward function R, and where $(\cdot)$ denotes the inner product. $Q^*_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a})$ is defined by a single vector $R_{\mathbf{a}}$ and therefore trivially PWLC.

The induction hypothesis is that for some time step $t+1$ we can represent the Q_{BG} value function as the maximum of the inner product of a belief and a set of vectors $\mathcal{V}^{t+1}_{\mathbf{a}}$ associated with joint action $\mathbf{a}$.

$$\forall_{b^{\vec{\theta}^{t+1}}} \quad Q^*_{\mathrm{B}}(b^{\vec{\theta}^{t+1}}, \mathbf{a}) = \max_{v^{t+1}_{\mathbf{a}} \in \mathcal{V}^{t+1}_{\mathbf{a}}} b^{\vec{\theta}^{t+1}} \cdot v^{t+1}_{\mathbf{a}}. \tag{3.7}$$

Now we have to prove that, given the induction hypothesis, Q_{BG} is also PWLC for t. I.e., we have to prove:

$$\forall_{b^{\vec{\theta}^t}} \quad Q^*_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a}) = \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}}} b^{\vec{\theta}^t} \cdot v^t_{\mathbf{a}}. \tag{3.8}$$

This is shown by picking up an arbitrary $b^{\vec{\theta}^t}$, for which the value of joint action $\mathbf{a}$ is given by (3.1), which we can rewrite as follows:

$$\begin{aligned} Q^*_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a}) &= R(b^{\vec{\theta}^t}, \mathbf{a}) + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) Q^*_{\mathrm{B}}(b^{\vec{\theta}^{t+1}}, \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1})) & (3.9) \\ &= b^{\vec{\theta}^t} \cdot R_{\mathbf{a}} + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} b^{\vec{\theta}^{t+1}} \cdot v^{t+1}_{\mathbf{a}'} & (3.10) \end{aligned}$$

where $R_{\mathbf{a}}$ is the immediate reward vector for joint action $\mathbf{a}$. In the second part $b^{\vec{\theta}^{t+1}}$ is the belief resulting from $b^{\vec{\theta}^t}$ by $\mathbf{a}$ and $\mathbf{o}^{t+1}$ and is given by:

$$\forall_{s^{t+1}} \qquad b^{\vec{\theta}^{t+1}}(s^{t+1}) = \frac{P(s^{t+1}, \mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a})}{P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a})}, \tag{3.11}$$

with

$$P(s^{t+1}, \mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) = \sum_{s^t} P(\mathbf{o}^{t+1} | \mathbf{a}, s^{t+1}) P(s^{t+1} | s^t, \mathbf{a}) b^{\vec{\theta}^t}(s^t). \tag{3.12}$$

Therefore we can write the second part of (3.10) as

$$\max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} \sum_{s^{t+1} \in \mathcal{S}} b^{\vec{\theta}^{t+1}}(s^{t+1}) v^{t+1}_{\mathbf{a}'}(s^{t+1}) =$$

$$\max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} \sum_{s^{t+1} \in \mathcal{S}} \left[\frac{P(s^{t+1}, \mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a})}{P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a})} \right] v^{t+1}_{\mathbf{a}'}(s^{t+1}) =$$

$$\max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} \sum_{s^{t+1} \in \mathcal{S}} P(s^{t+1}, \mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) v^{t+1}_{\mathbf{a}'}(s^{t+1}) =$$

$$\max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} \sum_{s^{t+1} \in \mathcal{S}} \left[\sum_{s^t} P(\mathbf{o}^{t+1} | \mathbf{a}, s^{t+1}) P(s^{t+1} | s^t, \mathbf{a}) b^{\vec{\theta}^t}(s^t) \right] v^{t+1}_{\mathbf{a}'}(s^{t+1}) =$$

$$\max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} \max_{\substack{v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \text{ s.t.} \\ \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'}} \sum_{s^t} \left[\sum_{s^{t+1} \in \mathcal{S}} P(\mathbf{o}^{t+1} | \mathbf{a}, s^{t+1}) P(s^{t+1} | s^t, \mathbf{a}) v^{t+1}_{\mathbf{a}'}(s^{t+1}) \right] b^{\vec{\theta}^t}(s^t). \tag{3.13}$$

Note that for a particular $\mathbf{a}, \mathbf{o}^{t+1}$ and $v_{\mathbf{a}'}^{t+1} \in \mathcal{V}_{\mathbf{a}'}^{t+1}$ we can define a function:

$$g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}}(s^t) = \sum_{s^{t+1} \in \mathcal{S}} P(\mathbf{o}|\mathbf{a}, s^{t+1}) P(s^{t+1}|s^t, \mathbf{a}) v_{\mathbf{a}'}^{t+1}(s^{t+1}). \tag{3.14}$$

This function defines a *gamma-vector* $g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}}$. For a particular $\mathbf{a}, \mathbf{o}^{t+1}$ we can define the set of gamma vectors that are consistent with a BG-policy $\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}$ for time step $t+1$ as

$$\mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \equiv \left\{ g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}} \mid v_{\mathbf{a}'}^{t+1} \in \mathcal{V}_{\mathbf{a}'}^{t+1} \wedge \beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}' \right\}. \tag{3.15}$$

Combining the gamma vector definition with (3.10) and (3.13) yields

$$Q_{\mathrm{B}}^*(b^{\vec{\theta}^t}, \mathbf{a}) = b^{\vec{\theta}^t} \cdot R_{\mathbf{a}} + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} \max_{g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}} \sum_{s^t} g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}}(s^t) b^{\vec{\theta}^t}(s^t). \tag{3.16}$$

Now let $g^*_{b^{\vec{\theta}^t},\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}$ denote the maximizing gamma-vector, i.e.:

$$\forall_{\mathbf{o}} \quad g^*_{b^{\vec{\theta}^t},\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \equiv \operatorname*{arg\,max}_{g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}} \sum_{s^t} g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}}(s^t) b^{\vec{\theta}^t}(s^t). \tag{3.17}$$

This allows to rewrite (3.16) to:

$$\begin{aligned} Q_{\mathrm{B}}^*(b^{\vec{\theta}^t}, \mathbf{a}) &= b^{\vec{\theta}^t} \cdot R_{\mathbf{a}} + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{\mathbf{o}^{t+1}} \sum_{s^t} g^*_{b^{\vec{\theta}^t},\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}(s^t) b^{\vec{\theta}^t}(s^t) \\ &= b^{\vec{\theta}^t} \cdot R_{\mathbf{a}} + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{s^t} \left[\sum_{\mathbf{o}^{t+1}} g^*_{b^{\vec{\theta}^t},\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}(s^t) \right] b^{\vec{\theta}^t}(s^t). \end{aligned} \tag{3.18}$$

The vectors for the different possible joint observations are now combined:

$$g^*_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}(s^t) \equiv \sum_{\mathbf{o}^{t+1}} g^*_{b^{\vec{\theta}^t},\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}(s^t), \tag{3.19}$$

which allows us to rewrite (3.18) as follows:

$$\begin{aligned} Q_{\mathrm{B}}^*(b^{\vec{\theta}^t}, \mathbf{a}) &= b^{\vec{\theta}^t} \cdot R_{\mathbf{a}} + \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \sum_{s^t} g^*_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}(s^t) b^{\vec{\theta}^t}(s^t) \\ &= \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \left(R_{\mathbf{a}} + g^*_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \right) \cdot b^{\vec{\theta}^t} &(3.20) \\ &= \max_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} v^{*,t}_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \cdot b^{\vec{\theta}^t} &(3.21) \end{aligned}$$

with

$$v^{*,t}_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} = R_{\mathbf{a}} + \sum_{\mathbf{o}^{t+1}} \left[\operatorname*{arg\,max}_{g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}}} \sum_{s^t} g_{\mathbf{a},\mathbf{o}}^{v_{\mathbf{a}'}^{t+1}}(s^t) b^{\vec{\theta}^t}(s^t) \right] \tag{3.22}$$

By defining

$$\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}} \equiv \left\{ v^{*,t}_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \mid \forall_{\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}} \right\}, \tag{3.23}$$

we can write

$$Q^*_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a}) = \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}, b^{\vec{\theta}^t}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t}, \tag{3.24}$$

which almost is what had to be proven. Although, for each $b^{\vec{\theta}^t}$, the set $\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}}$ can contain different vectors. However, it is clear that

$$\forall_{b^{\vec{\theta}^t}} \qquad \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}, b^{\vec{\theta}^t}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t} = \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t} \tag{3.25}$$

where

$$\mathcal{V}^t_{\mathbf{a}} \equiv \bigcup_{b^{\vec{\theta}^t} \in \mathcal{P}(\mathcal{S})} \mathcal{V}^t_{\mathbf{a}, b^{\vec{\theta}^t}}. \tag{3.26}$$

I.e., there is no vector in a different set $\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^{t\prime}}}$ that yields a higher value at $b^{\vec{\theta}^t}$ than the maximizing vector in $\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}}$. This can be easily seen as $\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}}$ is defined as the maximizing set of vectors at each belief point, and the different sets $\mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}}$ are all constructed using the same next time step policies and vectors, i.e., $v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'}$ s.t. $\beta_{\langle \vec{\theta}^t, \mathbf{a} \rangle}(\mathbf{o}^{t+1}) = \mathbf{a}'$ are the same. For a more formal proof see appendix A.

As a result we can write

$$Q^*_{\mathrm{B}}(b^{\vec{\theta}^t}, \mathbf{a}) = \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t}, \tag{3.27}$$

which is what had to be proven for $b^{\vec{\theta}^t}$. Realizing that we took no special assumption on $b^{\vec{\theta}^t}$, we can conclude this holds for all joint beliefs. □

4 Infinite horizon $\mathrm{Q_{BG}}$

Here we discuss how $\mathrm{Q_{BG}}$ can be extended to the infinite horizon. A naive translation of (3.1) to the infinite horizon would be given by:

$$Q_{\mathrm{B}}(b^{\vec{\theta}}, \mathbf{a}) = R(b^{\vec{\theta}}, \mathbf{a}) + \gamma \max_{\beta_{\left\langle b^{\vec{\theta}}, \mathbf{a} \right\rangle}} \sum_{\mathbf{o}} P(\mathbf{o}|b^{\vec{\theta}}, \mathbf{a}) Q_{\mathrm{B}}(b^{(\vec{\theta}, \mathbf{a}, \mathbf{o})}, \beta_{\left\langle b^{\vec{\theta}}, \mathbf{a} \right\rangle}(\mathbf{o})). \tag{4.1}$$

However, in the infinite-horizon case, the length of the joint action-observation histories is infinite, the set of all joint action-observation histories is infinite and there generally is an infinite number of corresponding joint beliefs. This means that it is not possible to convert a $\mathrm{Q_{BG}}$ function over joint action-observation histories to one over joint beliefs for the infinite horizon.[4]

Rather, we define a backup operator H_{B} for the infinite horizon that is directly making use of joint beliefs:

$$H_{\mathrm{B}} Q_{\mathrm{B}}(b, \mathbf{a}) = R(b, \mathbf{a}) + \gamma \max_{\beta_{\langle b, \mathbf{a} \rangle}} \sum_{\mathbf{o}} P(\mathbf{o}|b, \mathbf{a}) Q_{\mathrm{B}}(b^{\mathbf{ao}}, \beta_{\langle b, \mathbf{a} \rangle}(\mathbf{o})). \tag{4.2}$$

This is possible, because joint beliefs are still a sufficient statistic in the infinite-horizon case, as we will show next. After that, in section 4.2, we show that this backup operator is a contraction mapping.

[4] Also observe that the inductive proof of 3.1 does not hold in the infinite horizon case.

4.1 Sufficient statistic

The fact that Q^*_{BG} is a function over the joint belief space in the finite horizon case implies that a joint belief is a *sufficient statistic* of the history of the process. I.e., a joint belief contains enough information to uniquely predict the maximal achievable cumulative reward from this point on.

We will show that, also in the infinite-horizon case, a joint belief is a sufficient statistic for a Dec-POMDP with 1-step delayed communication. Let I^t denote the total information at some time step. Then we can write

$$I^t = \left(I^{t-1}, o^{t-1}_{\neq i}, \mathbf{a}^{t-1}, o^t_i\right), \tag{4.3}$$

with $I^0 = \left(b^0\right)$. I.e., the agent doesn't forget what he knew, he receives the observations of the other agents of the previous time step $o^{t-1}_{\neq i}$, and using this the agent is able to deduce $\mathbf{a}^{t-1}$, moreover he receives its own current observation. Effectively this means that $I^t = \left(b^0, \vec{\theta}^{t-1}, \mathbf{a}^{t-1}, o^t_i\right)$.

Now we want to show that rather than using $I^t = \left(b^0, \vec{\theta}^{t-1}, \mathbf{a}^{t-1}, o^t_i\right)$ we can also use $I^t_b = \left(b^{t-1}, \mathbf{a}^{t-1}, o^t_i\right)$, without lowering the obtainable value. Following [7], we notice that the belief update 3.2 implies that b^{t-1} is a sufficient statistic for the next joint belief b^t. Therefore, the rest of this proof focuses on showing that joint beliefs are also a sufficient statistic for the obtainable value.

When using I^t, an individual policy has the form $\pi^t_i : \vec{\Theta}^{t-1} \times \mathcal{A}^{t-1} \times \mathcal{O}_i \to \mathcal{A}_i$. Alternatively, we write such a policy as a set of policies for BGs $\pi^t_i = \left\{\beta_{\langle \vec{\theta}^{t-1}, \mathbf{a}\rangle, i}\right\}_{\langle \vec{\theta}^{t-1}, \mathbf{a}\rangle}$ where $\beta_{\langle \vec{\theta}^{t-1}, \mathbf{a}\rangle, i} : \mathcal{O}_i \to \mathcal{A}_i$. When we write π^* for the optimal joint policy with such a form, the expected optimal payoff of a particular time step t is given by:

$$E_{\pi^*}\left\{R(t)\right\} = \sum_{\vec{\theta}^{t-1}} \underbrace{\left[\sum_{\mathbf{o}^t}\left[\sum_s R(s, \beta^*_{\langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1}\rangle}(\mathbf{o}^t)) P(s|\vec{\theta}^t)\right] P(\mathbf{o}^t|\vec{\theta}^{t-1}, \mathbf{a}^{t-1})\right]}_{\text{Expectation of the BG for } \langle \vec{\theta}^{t-1}, \mathbf{a}^{t-1}\rangle} P(\vec{\theta}^{t-1}). \tag{4.4}$$

When using $I^t_b = \left(b^{t-1}, \mathbf{a}^{t-1}, o^t_i\right)$ as a statistic, the form of policies becomes $\pi^t_{b,i} : \mathcal{B} \times \mathcal{A}^{t-1} \times \mathcal{O}_i \to \mathcal{A}_i$, where $\mathcal{B} = \mathcal{P}(\mathcal{S})$ is the set of possible joint beliefs. Again, we also write $\beta_{\langle b^{t-1}, \mathbf{a}\rangle, i}$.

Now, we need to show that for all t':

$$v^{t'}(I^{t'}) = E_{\pi^*}\left\{\sum_{t=t'}^{\infty} \gamma^{t-t'} R(t)\right\} = E_{\pi^*_b}\left\{\sum_{t=t'}^{\infty} \gamma^{t-t'} R(t)\right\} = v^{t'}(I^{t'}_b) \tag{4.5}$$

Note that

$$E_{\pi^*}\left\{\sum_{t=t'}^{\infty} \gamma^{t-t'} R(t)\right\} = \sum_{t=t'}^{\infty} \gamma^{t-t'} E_{\pi^*}\left\{R(t)\right\} \tag{4.6}$$

and similar for π^*_b. Therefore we only need to show that

$$\forall_{t=0,1,2,\ldots} \quad E_{\pi^*}\left\{R(t)\right\} = E_{\pi^*_b}\left\{R(t)\right\}. \tag{4.7}$$

If we assume that for an arbitrary time step $t-1$ the different possible joint beliefs b^{t-1} corresponding to all $\vec{\theta}^{t-1} \in \vec{\Theta}^{t-1}$ are a sufficient statistic for the expected reward for time steps $0, ..., t-1$, we can write:

$$E_{\pi^*_b}\left\{R(t)\right\} = \sum_{b^{t-1}} \underbrace{\left[\sum_{\mathbf{o}^t}\left[\sum_s R(s, \beta^*_{\langle b^{t-1}, \mathbf{a}^{t-1}\rangle}(\mathbf{o}^t)) b^t_{\mathbf{ao}}(s)\right] P(\mathbf{o}^t|b^t_{\mathbf{ao}}(s), \mathbf{a}^{t-1})\right]}_{\text{Expectation of the BG for } \langle b^{t-1}, \mathbf{a}^{t-1}\rangle} P(b^{t-1}). \tag{4.8}$$

Because $P(s|\vec{\theta}^t) \equiv P(s|b^{\vec{\theta}^t}) = b^t_{\mathbf{ao}}(s)$, where $b^t_{\mathbf{ao}}(s)$ is the belief resulting from $b^{\vec{\theta}^{t-1}}$ via $\mathbf{a}, \mathbf{o}$, and $P(\mathbf{o}^t|\vec{\theta}^{t-1}, \mathbf{a}^{t-1}) \equiv P(\mathbf{o}^t|b^{\vec{\theta}^{t-1}}, \mathbf{a}^{t-1})$, we can conclude that also for this time step $E_{\pi^*}\{R(t)\} = E_{\pi^*_b}\{R(t)\}$, meaning that maintaining joint beliefs is a sufficient statistic for time step t as well. A base case is given at time step 0, because $I^0 = I^0_b = (b^0)$. By induction it follows that joint beliefs are a sufficient statistic for all time steps.

4.2 Contraction mapping

To improve the readability of the formulas, in this section Q_{B} is written as simply Q.

Theorem 4.1 *The infinite-horizon Q_{BG}-backup operator* (4.2) *is a contraction mapping under the following supreme norm:*

$$\left\|Q - Q'\right\| = \sup_b \max_{\mathbf{a}} \left| \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \; \left[Q(b^{\mathbf{ao}}, \beta_{max}(Q)(\mathbf{o})) - Q'(b^{\mathbf{ao}}, \beta_{max}(Q')(\mathbf{o}))\right] \right|, \tag{4.9}$$

where

$$\beta_{max}(Q) = \arg\max_{\beta_{\langle b,\mathbf{a}\rangle}} \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) Q(b^{\mathbf{ao}}, \beta_{\langle b,\mathbf{a}\rangle}(\mathbf{o})) \tag{4.10}$$

is the maximizing BG policy according to Q.

Proof We have to prove that

$$\left\|H_{\mathrm{B}}Q - H_{\mathrm{B}}Q'\right\| \le \gamma \left\|Q - Q'\right\|. \tag{4.11}$$

When applying the backup we get:

$$\begin{aligned} \left\|H_{\mathrm{B}}Q - H_{\mathrm{B}}Q'\right\| &= \sup_b \max_{\mathbf{a}} \left| \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \; \left[H_{\mathrm{B}}Q(b^{\mathbf{ao}}, \beta_{\max}(Q)(\mathbf{o})) - H_{\mathrm{B}}Q'(b^{\mathbf{ao}}, \beta_{\max}(Q')(\mathbf{o}))\right] \right| \\ &= \sup_b \max_{\mathbf{a}} \left| \left[\sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) H_{\mathrm{B}}Q(b^{\mathbf{ao}}, \beta_{\max}(Q)(\mathbf{o})) \right] \right. \\ &\qquad \left. - \left[\sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) H_{\mathrm{B}}Q'(b^{\mathbf{ao}}, \beta_{\max}(Q')(\mathbf{o})) \right] \right|. \end{aligned} \tag{4.12}$$

When, without loss of generality, we assume that b, $\mathbf{a}$ are the maximizing arguments, and if we assume that the first part (the summation over HQ) is larger then the second part (that over HQ'), we can write

$$\left\|H_{\mathrm{B}}Q - H_{\mathrm{B}}Q'\right\| = \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \left[H_{\mathrm{B}}Q(b^{\mathbf{ao}}, \beta_{\max}(Q)(\mathbf{o})) - H_{\mathrm{B}}Q'(b^{\mathbf{ao}}, \beta_{\max}(Q')(\mathbf{o}))\right] \tag{4.13}$$

If we use $\beta_{\max}(Q)$ instead of $\beta_{\max}(Q')$ in the last term, we are subtracting less, so we can write

$$\left\|H_{\mathrm{B}}Q - H_{\mathrm{B}}Q'\right\| \le \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \left[H_{\mathrm{B}}Q(b^{\mathbf{ao}}, \beta_{\max}(Q)(\mathbf{o})) - H_{\mathrm{B}}Q'(b^{\mathbf{ao}}, \beta_{\max}(Q)(\mathbf{o}))\right] \tag{4.14}$$

Now let $\beta_{\max}(Q)(\mathbf{o}) = \mathbf{a}'$, then we get

$$
\begin{aligned}
&= \gamma \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \sum_{\mathbf{o}'} P(\mathbf{o}'|b^{\mathbf{ao}},\mathbf{a}') \left[Q(b^{\mathbf{aoa'o'}}, \beta_{\max}(Q)(\mathbf{o}')) - Q'(b^{\mathbf{aoa'o'}}, \beta_{\max}(Q')(\mathbf{o}')) \right] \\
&\leq \gamma \sum_{\mathbf{o}} P(\mathbf{o}|b,\mathbf{a}) \sup_{b'} \max_{\mathbf{a}'} \left| \sum_{\mathbf{o}'} P(\mathbf{o}'|b',\mathbf{a}') \left[Q(b'^{\mathbf{a'o'}}, \beta_{\max}(Q)(\mathbf{o}')) - Q'(b'^{\mathbf{a'o'}}, \beta_{\max}(Q')(\mathbf{o}')) \right] \right| \\
&= \gamma \sup_{b'} \max_{\mathbf{a}'} \left| \sum_{\mathbf{o}'} P(\mathbf{o}'|b',\mathbf{a}') \left[Q(b'^{\mathbf{a'o'}}, \beta_{\max}(Q)(\mathbf{o}')) - Q'(b'^{\mathbf{a'o'}}, \beta_{\max}(Q')(\mathbf{o}')) \right] \right| \\
&= \gamma \left\| Q - Q' \right\|
\end{aligned}
\tag{4.15}
$$

For $\gamma \in (0,1)$ this is a contraction mapping. □

4.3 Infinite horizon $\mathrm{Q_{BG}}$

The fact that (4.2) is a contraction mapping means that there is a fixed point, which is the optimal infinite horizon $\mathrm{Q_{BG}}$-value function $Q_{\mathrm{B}}^{*,\infty}(b,\mathbf{a})$ [2]. Together with the fact that Q_{B}^{*} for the finite horizon is PWLC, this means we can approximate $Q_{\mathrm{B}}^{*,\infty}(b,\mathbf{a})$ with arbitrary accuracy using a PWLC value function.

5 The optimal Dec-POMDP value function Q^*

Here we show that it is not possible to convert the optimal Dec-POMDP Q-value function, $Q^*(\vec{\theta}^t,\mathbf{a})$, to $Q^*(b^{\vec{\theta}^t},\mathbf{a})$ a similar function over joint beliefs.

Lemma 5.1 *The optimal Q^* value function for a Dec-POMDP, given by:*

$$
Q^*(\vec{\theta}^t,\mathbf{a}) = R(\vec{\theta}^t,\mathbf{a}) + \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|\vec{\theta}^t,\mathbf{a}) Q^*(\vec{\theta}^{t+1}, \pi^*(\vec{\theta}^{t+1})). \tag{5.1}
$$

generally is not a function over the belief space.

Proof If Q^* would be a function over the belief space, as in section 3.1, it should hold that it is not possible that different joint action-observation histories specify different values, while the underlying joint belief is the same. Following the same argumentation as in section 3.1, it should hold that if $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$, it holds that

$$
\sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|\vec{\theta}^{t,a},\mathbf{a}) Q^*(\vec{\theta}^{t+1,a}, \pi^*(\vec{\theta}^{t+1,a})) = \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1}|\vec{\theta}^{t,b},\mathbf{a}) Q^*(\vec{\theta}^{t+1,b}, \pi^*(\vec{\theta}^{t+1,b})), \tag{5.2}
$$

given that $b^{\vec{\theta}^{t+1,a}} = b^{\vec{\theta}^{t+1,b}}$ implies $Q^*(\vec{\theta}^{t+1,a},\mathbf{a}) = Q^*(\vec{\theta}^{t+1,b},\mathbf{a})$. Again, the observation probabilities, resulting joint beliefs and thus $Q^*(\vec{\theta}^{t+1},\mathbf{a})$-values are equal. However now, it might be possible that the optimal policy π^* specifies different actions at the next time step which would lead to different future rewards. I.e., for Q^* to be convertible to a function over joint beliefs,

$$
\forall_{\mathbf{o}^{t+1}} \quad \pi^*(\vec{\theta}^{t+1,a}) = \pi^*(\vec{\theta}^{t+1,b}) \tag{5.3}
$$

should hold if $b^{\vec{\theta}^{t,a}} = b^{\vec{\theta}^{t,b}}$. This, however, is not provable and we will provide a counter example using the the horizon 3 dec-tiger problem [4] here. The observations are denoted L=hear tiger left and R=hear tiger right, the actions are written Li=listen, OL=open left and OR=open right.

Consider the following two joint action-observation histories for time step $t = 1$: $\vec{\theta}^{1,a} = \langle (Li, L), (Li, R) \rangle$ and $\vec{\theta}^{1,b} = \langle (Li, R), (Li, L) \rangle$. For these histories we $b^{\vec{\theta}^{1,a}} = b^{\vec{\theta}^{1,b}} = \langle 0.5, 0.5 \rangle$. Now we consider the future reward for $\mathbf{a} = \langle Li, Li \rangle$ and $\mathbf{o} = \langle L, R \rangle$. For this case, the observation probabilities are equal $P(\langle L, R \rangle | \vec{\theta}^{1,a}, Li) = P(\langle L, R \rangle | \vec{\theta}^{1,b}, Li)$ and the successor joint action-observation histories $\vec{\theta}^{2,a} = \langle (Li, L, Li, L), (Li, R, Li, R) \rangle$ and $\vec{\theta}^{2,b} = \langle (Li, R, Li, L), (Li, L, Li, R) \rangle$ both specify the same joint belief: $b^{\vec{\theta}^{2,a}} = b^{\vec{\theta}^{2,b}} = \langle 0.5, 0.5 \rangle$. However,

$$\pi^*(\vec{\theta}^{2,a}) = \langle OL, OR \rangle \neq \langle Li, Li \rangle = \pi^*(\vec{\theta}^{2,b}). \tag{5.4}$$

So even though the induction hypothesis says that

$$\forall_{\mathbf{a}} \quad Q^*(\vec{\theta}^{t+1,a}, \mathbf{a}) = Q^*(\vec{\theta}^{t+1,b}, \mathbf{a}), \tag{5.5}$$

different actions may be selected by π^* for $\vec{\theta}^{t+1,a}$ and $\vec{\theta}^{t+2,a}$ and therefore (5.3) and thus (5.2) are not guaranteed to hold. □

References

[1] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Math. Oper. Res.*, 27(4):819–840, 2002.

[2] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 2nd edition, 2001.

[3] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 136–143, Washington, DC, USA, 2004. IEEE Computer Society.

[4] Ranjit Nair, Milind Tambe, Makoto Yokoo, David V. Pynadath, and Stacy Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 705–711, 2003.

[5] Frans A. Oliehoek and Nikos Vlassis. Q-value functions for decentralized POMDPs. In *Proc. of Int. Joint Conference on Autonomous Agents and Multi Agent Systems*, May 2007. to appear.

[6] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory.* The MIT Press, July 1994.

[7] Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, September 1973.

A Sub-proof of PWLC property

We have to show that the maximizing vector given b is the maximizing vector at b, i.e., that the following holds:

$$\forall_{b^{\vec{\theta}^t}} \quad \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}, b^{\vec{\theta}^t}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t} = \max_{v^t_{\mathbf{a}} \in \bigcup_{b^{\vec{\theta}^t} \in \mathcal{P}(\mathcal{S})} \mathcal{V}^t_{\mathbf{a}, b^{\vec{\theta}^t}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t}.$$

Proof (By contradiction): For an arbitrary $b^{\vec{\theta}^t}$, suppose there is a different joint belief $b^{\vec{\theta}^{t\prime}}$ such that

$$\max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^t}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t} < \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a},b^{\vec{\theta}^{t\prime}}}} v^t_{\mathbf{a}} \cdot b^{\vec{\theta}^t}.$$

According to (3.23) and (3.21), this would mean that

$$\max_{\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} v^{*,t}_{b^{\vec{\theta}^t},\mathbf{a},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} \cdot b^{\vec{\theta}^t} < \max_{\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}} v^{*,t}_{b^{\vec{\theta}^{t\prime}},\mathbf{a},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}} \cdot b^{\vec{\theta}^t}$$

which implies that:

$$\max_{\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} \left(R_{\mathbf{a}} + \sum_{\mathbf{o}^{t+1}} \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}}} \sum_{s^t} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}}(s^t) b^{\vec{\theta}^t}(s^t) \right] \right) \cdot b^{\vec{\theta}^t}$$
$$< \max_{\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}} \left(R_{\mathbf{a}} + \sum_{\mathbf{o}^{t+1}} \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}}} \sum_{s^t} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}}(s^t) b^{\vec{\theta}^{t\prime}}(s^t) \right] \right) \cdot b^{\vec{\theta}^t}$$

Because $R_{\mathbf{a}}$ is the same for both vectors, this means that

$$\max_{\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} \left(\sum_{\mathbf{o}^{t+1}} \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^t} \right] \right) \cdot b^{\vec{\theta}^t} < \max_{\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}} \left(\sum_{\mathbf{o}^{t+1}} \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^{t\prime}} \right] \right) \cdot b^{\vec{\theta}^t}$$

thus:

$$\max_{\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} \sum_{\mathbf{o}^{t+1}} \left(\left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^t} \right] \cdot b^{\vec{\theta}^t} \right) < \max_{\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}} \sum_{\mathbf{o}^{t+1}} \left(\left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^{t\prime}} \right] \cdot b^{\vec{\theta}^t} \right), \tag{A.1}$$

would have to hold. However, because the possible choices for $\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}$ and $\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}$ are identical, we know that $\mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}} = \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}}$, and therefore that

$$\forall_{\mathbf{o}} \quad \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^t,\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^t} \right] \cdot b^{\vec{\theta}^t} \geq \left[\arg\max_{g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \in \mathcal{G}_{\mathbf{a},\mathbf{o},\beta_{\langle\vec{\theta}^{t\prime},\mathbf{a}\rangle}}} g^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{a},\mathbf{o}} \cdot b^{\vec{\theta}^{t\prime}} \right] \cdot b^{\vec{\theta}^t},$$

contradicting (A.1). □

Acknowledgements

The research reported here is supported by the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024. This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS_Conhecimento Program that includes FEDER funds.

IAS reports

This report is in the series of IAS technical reports. The series editor is Stephan ten Hagen (stephanh@science.uva.nl). Within this series the following titles appeared:

P.J. Withagen and F.C.A. Groen and K. Schutte *Shadow detection using a physical basis.* Technical Report IAS-UVA-07-02, Informatics Institute, University of Amsterdam, The Netherlands, January 2007.

P.J. Withagen and K. Schutte and F.C.A. Groen *Global intensity correction in dynamic scenes.* Technical Report IAS-UVA-07-01, Informatics Institute, University of Amsterdam, The Netherlands, January 2007.

A. Noulas and B. Kröse *Probabilistic audio visual sensor fusion for speaker detection.* Technical Report IAS-UVA-06-04, Informatics Institute, University of Amsterdam, The Netherlands, March 2006.

All IAS technical reports are available for download at the IAS website, http://www.science.uva.nl/research/ias/publications/reports/.