Chapter 3

Optimal Value Functions for Dec-POMDPs

Over the last half century, the single-agent MDP framework has received much attention, and many results are known (Bellman, 1957b,a; Howard, 1960; Puterman, 1994; Sutton and Barto, 1998; Bertsekas, 2007). In particular it is known that an optimal plan, or policy, can be extracted from the optimal action-value, or Q-value, function $Q^*(s,a)$, and that the latter can be calculated efficiently. For POMDPs, similar results are available, although finding an optimal solution is harder (PSPACE-complete for finite-horizon problems, Papadimitriou and Tsitsiklis, 1987).

This chapter discusses how value functions can be computed for decentralized settings, and how policies can be extracted from those value functions. In particular, it treats the value functions that result of superimposing different assumptions with respect to communication on top of the Dec-POMDP model. The common assumption is that this communication is noise and cost free, the differences are with respect to the delay this communication incurs. We consider:

- The plain Dec-POMDP setting which assumes no communication. This can be interpreted as a delay of at least $h$ stages.
- Instantaneous communication.
- One-step delayed communication.
- $k$-steps delayed communication.

The main contribution of this chapter lies in the contribution to the theory of Dec-POMDPs. In particular, it addresses the previously outstanding issue whether Q-value functions can be defined for Dec-POMDPs just as in (PO)MDPs, and whether an optimal policy can be extracted from such Q-value functions. Most algorithms for planning in Dec-POMDPs are based on some version of policy search, and a proper theory of Q-value functions in Dec-POMDPs has been lacking. Given
the wide range of applications of value functions in single-agent decision-theoretic planning, we expect that such a theory for Dec-POMDPs can provide great benefits, both in terms of providing insight as well as guiding the design of solution algorithms.

A second contribution is the integrated overview of value-functions for decentralized settings. For the communicative settings already some results are known. In particular the immediate communication case reduces to the POMDP setting, for which many results are known, and Dec-MDP settings with delayed communication have received some attention in control theory literature. Yet, to the author’s knowledge, there has never been an overview presenting these different settings in an integrated manner. Also, this chapter extends to the Dec-POMDP setting with delayed communication.

3.1 No Communication

This section shows how value functions can be defined for Dec-POMDPs of a finite horizon $h$. That is we assume that the agents cannot communicate or, equivalently, that the delay of communication is $h$. These value functions can be employed by representing a Dec-POMDP as a series of Bayesian Games, which were discussed in Subsection 2.1.1.2. This idea of using a series of BGs to find policies for a Dec-POMDP was proposed in an approximate setting by Emery-Montemerlo et al. (2004). In particular, they showed that using series of BGs and an approximate payoff function, they were able to obtain approximate solutions for Dec-POMDPs.

The main result of this section is that an optimal joint Dec-POMDP policy can be computed from the solution of a sequence of Bayesian games. That is, there exists at least one Q-value function $Q^*_\pi$ (corresponding to an optimal policy $\pi^*$) that, when used as the payoff function of a sequence of Bayesian games, will yield an optimal solution. Thus, the results of Emery-Montemerlo et al. (2004) are extended to include the optimal setting.

Subsection 3.1.4 shows that without assuming a particular optimal joint policy $\pi^*$, it is not possible to compute an optimal value function of the form $Q^*_t(\vec{\theta}^t, a)$. Rather, as Subsection 3.1.5 demonstrates, the optimal $Q^*_t$ for a stage $t$ depends on the past joint policy executed over stages $0,\ldots,t-1$.

3.1.1 Modeling Dec-POMDPs as Series of Bayesian Games

Bayesian games can be used to model Dec-POMDPs. Essentially, a Dec-POMDP can be seen as a tree where nodes are joint action-observation histories and edges represent joint actions and observations, as illustrated in Figure 3.1. At a specific stage $t$ in a Dec-POMDP, the main difficulty in coordinating action selection is presented by the fact that each agent has its own individual action-observation history. That is, there is no global signal that the agents can use to coordinate their actions. This situation can be conveniently modeled by a Bayesian game as we will now discuss.

At a time step $t$, one can directly associate the primitives of a Dec-POMDP with those of a BG with identical payoffs: the actions of the agents are the same in
Figure 3.1: A Dec-POMDP can be seen as a tree of joint actions and observations. The indicated planes correspond with the Bayesian games for the first two stages.

both cases, and the types of agent $i$ correspond to its action-observation histories $\Theta_i \equiv \vec{\Theta}_i^t$. Figure 3.2 shows the Bayesian games for $t = 0$ and $t = 1$ for a fictitious Dec-POMDP with 2 agents.

We denote the payoff function of the BG that models a stage of a Dec-POMDP by $Q(\vec{\theta}^t, a)$. This payoff function should be naturally defined in accordance with the value function of the planning task. For instance, Emery-Montemerlo et al. define $Q(\vec{\theta}^t, a)$ as the $Q_{\text{MDP}}$-value of the underlying MDP. We will more extensively discuss the payoff function in Subsection 3.1.2.

The probability $\Pr(\theta)$ is equal to the probability of the joint action-observation history $\vec{\theta}^t$ to which $\theta$ corresponds and depends on the past joint policy $\varphi^t = (\delta^0, \ldots, \delta^{t-1})$ and the initial state distribution. It can be calculated as the marginal of (2.5.6):

$$
\Pr(\theta) = \Pr(\vec{\theta}^t|\varphi^t, b^0) = \sum_{s^t \in S} \Pr(s^t, \vec{\theta}^t|\varphi^t, b^0).
$$

When only considering pure joint policies $\varphi^t$, the action probability component $\pi(a|\vec{\theta})$ in (2.5.6) (which corresponds to $\varphi(a|\vec{\theta})$ here) is 1 for joint action-observation histories $\vec{\theta}^t$ that are ‘consistent’ with the past joint policy $\varphi^t$ and 0 otherwise. We say that an action-observation $\vec{\theta}_i$ history is consistent with a pure policy $\pi_i$ if it can occur when executing $\pi_i$, i.e., when the actions in $\vec{\theta}_i$ would be selected by $\pi_i$. Let us more formally define this consistency as follows.
We will also write $\mathbf{\bar{\theta}}^t_i$ for the restriction of $\mathbf{\bar{\theta}}^t_i$ to stages $0, \ldots, k$ (with $0 \leq k < t$). An action-observation history $\mathbf{\bar{\theta}}^t_i$ of agent $i$ is consistent with a pure policy $\pi_i$ if and only if at each time step $k$ with $0 \leq k < t$

$$\pi_i(\mathbf{\bar{\theta}}^k_i) = \pi_i(\mathbf{\bar{a}}^k_i) = a^k_i$$

is the $(k + 1)$-th action in $\mathbf{\bar{\theta}}^t_i$. A joint action-observation history $\mathbf{\bar{\theta}}^t = (\mathbf{\bar{\theta}}^t_1, \ldots, \mathbf{\bar{\theta}}^t_n)$ is consistent with a pure joint policy $\pi = (\pi_1, \ldots, \pi_n)$ if each individual $\mathbf{\bar{\theta}}^t_i$ is consistent with the corresponding individual policy $\pi_i$. $C$ is the indicator function for consistency. For instance $C(\mathbf{\bar{\theta}}^t, \pi)$ ‘filters out’ the action-observation histories $\mathbf{\bar{\theta}}^t$ that are inconsistent with a joint pure policy $\pi$:

$$C(\mathbf{\bar{\theta}}^t, \pi) = \begin{cases} 1 & , \mathbf{\bar{\theta}}^t = (\mathbf{o}^0, \pi(\mathbf{o}^0), \mathbf{o}^1, \pi(\mathbf{o}^0, \mathbf{o}^1), \ldots) \\ 0 & , \text{otherwise.} \end{cases} \tag{3.1.2}$$

We will also write $\mathbf{\bar{\Theta}}^t_\pi = \{ \mathbf{\bar{\theta}}^t | C(\mathbf{\bar{\theta}}^t, \pi) = 1 \}$ for the set of $\mathbf{\bar{\theta}}^t$ consistent with $\pi$.

This definition allows us to write

$$\Pr(\mathbf{\bar{\theta}}^t | \mathbf{\phi}^t, \mathbf{b}^0) = C(\mathbf{\bar{\theta}}^t, \mathbf{\phi}^t) \sum_{s^t \in S} \Pr(s^t, \mathbf{\bar{\theta}}^t | \mathbf{b}^0) \tag{3.1.3}$$

with

$$\Pr(s^t, \mathbf{\bar{\theta}}^t | \mathbf{b}^0) = \sum_{s^{t-1} \in S} \Pr(o^t | a^{t-1}, s^t) \Pr(s^t | s^{t-1}, a^{t-1}) \Pr(s^{t-1}, \mathbf{\bar{\theta}}^{t-1} | \mathbf{b}^0). \tag{3.1.4}$$

Figure 3.2 illustrates how the indicator function ‘filters out’ action-observation histories: when $\pi^{t=0}(\mathbf{\bar{\theta}}^{t=0}) = \langle a_1, a_2 \rangle$, only the non-shaded part of the BG for $t = 1$ ‘can be reached’ (has positive probability).

### 3.1.2 The Q-Value Function of an Optimal Joint Policy

Given the perspective of a Dec-POMDP interpreted as a series of BGs as outlined in the previous section, the solution of the BG for stage $t$ is a joint decision rule...
\(\delta^t\) for that stage. If the payoff function for the BG is chosen well, the quality of \(\delta^t\) should be high. Emery-Montemerlo et al. (2004) try to find a good joint policy \(\pi = (\delta^0, \ldots, \delta^{h-1})\) by a procedure we refer to as forward-sweep policy computation (FSPC): in one sweep forward through time, the BG for each stage \(t = 0, 1, \ldots, h-1\) is consecutively solved. As such, the payoff functions for the BGs constitute a Q-value function for the Dec-POMDP.

Here, we show that there exists an optimal Q-value function \(Q^*:\) when using this as the payoff functions for the BGs, forward-sweep policy computation will lead to an optimal joint policy \(\pi^* = (\delta^0, *, \ldots, \delta^{h-1}, *)\). In particular, given an optimal joint policy \(\pi^*\) we identify a normative description of \(Q_{\pi^*}\) as the Q-value function for an optimal joint policy.

**Proposition 3.1** (Value of an optimal joint policy.). The expected cumulative reward over stages \(t, \ldots, h - 1\) induced by \(\pi^*\), an optimal pure joint policy for a Dec-POMDP, is given by:

\[
V^t(\pi^*) = \sum_{\bar{\theta}^t \in \bar{\Theta}^t_{\pi^*}} \Pr(\bar{\theta}^t|b^0)Q_{\pi^*}(\bar{\theta}^t, \pi^*(\bar{\theta}^t)),
\]

where \(\bar{\theta}^t = (\bar{o}^t, \bar{a}^t)\), where \(\pi^*(\bar{\theta}^t) = \pi^*(\bar{o}^t)\) denotes the joint action \(\pi^*\) specifies for \(\bar{o}^t\), and where

\[
Q_{\pi^*}(\bar{\theta}^t, a) = R(\bar{\theta}^t, a) + \sum_{o^t+1 \in O} \Pr(o^{t+1}|\bar{\theta}^t, a)Q_{\pi^*}(\bar{\theta}^{t+1}, \pi^*(\bar{\theta}^{t+1}))
\]

is the Q-value function for \(\pi^*\), which gives the expected cumulative future reward when taking joint action \(a\) at \(\bar{\theta}^t\) given that \(\pi^*\) is followed hereafter.

**Sketch of proof.** By filling out (2.5.5) for an optimal pure joint policy \(\pi^*\) it is possible to derive the equations. This derivation is listed in Appendix D. \(\square\)

### 3.1.3 Deriving an Optimal Joint Policy

At this point we have derived \(Q_{\pi^*}\), a Q-value function for an optimal joint policy. Now, we extend the results of Emery-Montemerlo et al. (2004) into the exact setting by the following theorem:

**Theorem 3.1** (Optimality of FSPC). Applying forward-sweep policy computation using \(Q_{\pi^*}\) as defined by (3.1.6) yields an optimal joint policy.

**Proof.** Note that, per definition, an optimal Dec-POMDP policy \(\pi^*\) maximizes the expected future reward \(V^t(\pi^*)\) specified by (3.1.5). Therefore \(\delta^t\), the optimal decision rule for stage \(t\), is identical to an optimal joint policy \(\beta^t\) for the Bayesian game for time step \(t\), if the payoff function of the BG is given by \(Q_{\pi^*}\), that is:

\[
\delta^t = \beta^t = \arg \max_{\beta^t} \sum_{\bar{\theta}^t \in \bar{\Theta}^t_{\pi^*}} \Pr(\bar{\theta}^t|b^0)Q_{\pi^*}(\bar{\theta}^t, \beta^t(\bar{\theta}^t)).
\]
Equation (3.1.7) tells us that $\delta^{t,*} \equiv \beta^{t,*}$. This means that it is possible to construct the complete optimal Dec-POMDP policy $\pi^* = (\delta^{0,*}, \ldots, \delta^{h-1,*})$, by computing $\delta^{t,*}$ for all $t$.

A subtlety in the calculation of $\pi^*$ is that (3.1.7) itself is dependent on an optimal joint policy, as the summation is over all $\tilde{\theta}^t \in \tilde{\Theta}^t_{\pi^*} \equiv \{ \tilde{\theta}^t \mid C(\tilde{\theta}^t, \pi^*) = 1 \}$. This is resolved by realizing that only the past actions influence which action-observation histories can be reached at time step $t$. If we denote the optimal past joint policy by $\varphi^{t,*} = (\delta^{0,*}, \ldots, \delta^{t-1,*})$, we have that $\tilde{\Theta}^t_{\pi^*} = \tilde{\Theta}^t_{\varphi^{t,*}}$, and therefore that:

$$
\beta^{t,*} = \arg \max_{\beta^t} \sum_{\tilde{\theta}^t \in \tilde{\Theta}^t_{\varphi^{t,*}}} \Pr(\tilde{\theta}^t | b^0) Q_{\pi^*}(\tilde{\theta}^t, \beta^t(\tilde{\theta}^t)).
$$

(3.1.8)

This can be solved in a forward manner for time steps $t = 0, 1, 2, \ldots, h - 1$, because at every time step $\varphi^{t,*}$ will be available: it is specified by $(\beta^{0,*}, \ldots, \beta^{t-1,*})$ the solutions of the previously solved BGs.

In this way, we have identified how a Q-value function of the form $Q_{\pi^*}(\tilde{\theta}^t, a^t)$ can be used in the planning phase to find an optimal policy through solution of a series of BGs. Note that the fact that the true AOH, $\tilde{\theta}^t$, is not observed during execution is not a problem: the joint policy is constructed in the (off-line) planning phase, based on the expectation over all possible AOHs. This expectation is implicit in what can be called the ‘BG operator’, i.e., the solution of a BG.

### 3.1.4 Computing an Optimal Q-Value Function

So far we discussed that $Q_{\pi^*}$ can be used to find an optimal joint policy $\pi^*$. Unfortunately, when an optimal joint policy $\pi^*$ is not known, computing $Q_{\pi^*}$ itself is impractical, as we will discuss here. This is in contrast with the (fully observable) single-agent case where the optimal Q-values can be found relatively easily in a single sweep backward through time.

In Section 3.1.2, the optimal expected return was expressed as $Q_{\pi^*}(\tilde{\theta}^t, a^t)$ by assuming an optimal joint policy $\pi^*$ is followed up to the current stage $t$. However, when no such previous policy is assumed, the optimal expected return is not defined.

**Proposition 3.2** (No $Q_{\pi^*}(\tilde{\theta}^t, a^t)$ except for last stage). For a pair $(\tilde{\theta}^t, a^t)$ with $t < h - 1$ the optimal value $Q_{\pi^*}(\tilde{\theta}^t, a^t)$ cannot be defined without assuming some (possibly randomized) past policy $\varphi^{t+1} = (\delta^{0,*}, \ldots, \delta^t)$. Only for the last stage $t = h - 1$ such expected reward is defined as

$$
Q_{\pi^*}(\tilde{\theta}^{h-1}, a^{h-1}) \equiv R(\tilde{\theta}^{h-1}, a^{h-1})
$$

(3.1.9)

without assuming a past policy.

**Proof.** Let us try to deduce $Q_{\pi^*}(\tilde{\theta}^t, a^t)$ the optimal value for a particular $\tilde{\theta}^t$ assuming the $Q_{\pi^*}$-values for the next time step $t + 1$ are known. The $Q_{\pi^*}(\tilde{\theta}^t, a^t)$-values for each of the possible joint actions can be evaluated as follows

$$
\forall a \quad Q_{\pi^*}(\tilde{\theta}^t, a^t) = R(\tilde{\theta}^t, a^t) + \sum_{\sigma^{t+1}} \Pr(o^{t+1} | \tilde{\theta}^t, a^t) Q_{\pi^*}(\tilde{\theta}^{t+1}, \sigma^{t+1,*}(\tilde{\theta}^{t+1})).
$$

(3.1.10)
where $\delta^{t+1,*}$ is an optimal decision rule for the next stage. But what should $\delta^{t+1,*}$ be? If we assume that up to stage $t+1$ we followed a particular (possibly randomized) past joint policy $\varphi^{t+1}$, then

$$\delta^{t+1,*}_\varphi = \arg \max_{\delta^{t+1}} \sum_{\tilde{\theta}^{t+1} \in \tilde{\Theta}^{t+1}} \Pr(\tilde{\theta}^{t+1}|\varphi^{t+1},b^0)Q^*(\tilde{\theta}^{t+1},\beta^{t+1}(\tilde{\theta}^{t+1})). \quad (3.1.11)$$

is optimal. However, there are many pure and infinite randomized past policies $\varphi^{t+1}$ that are consistent with $\tilde{\theta}^t,a^t$, and thus many different $\tilde{\Theta}^{t+1}_\varphi$ over which the above maximization could take place, leading to many $\delta^{t+1,*}_\varphi$ that might be optimal. The conclusion we can draw is that $Q^*(\tilde{\theta}^t,a^t)$ is ill-defined without $\Pr(\tilde{\theta}^{t+1}|\varphi^{t+1},b^0)$, the probability distribution (belief) over joint action-observation histories, which is induced by $\varphi^{t+1}$, the policy followed for stages $0,\ldots,t$. \(\square\)

Let us investigate what the consequences of this insight are for the formulation of the optimal Q-value function as defined in Section 3.1.2. Consider $\pi^*(\tilde{\theta}^{t+1})$ in (3.1.6). This optimal policy is a mapping from observation histories to actions $\pi^*: \tilde{O} \rightarrow A$ induced by the individual policies and observation histories. This means that for two joint action-observation histories with the same joint observation history, $\pi^*$ results in the same joint action. That is $\forall \tilde{a},\tilde{o},\tilde{a}', \pi^*(\langle \tilde{a},\tilde{o} \rangle) = \pi^*(\langle \tilde{a}',\tilde{o} \rangle)$. Effectively this means that when we reach some $\tilde{\theta}^t \notin \tilde{\Theta}^t_{\pi^*}$, say through a mistake, $\pi^*$ continues to specify actions as if no mistake ever happened: That is, still assuming that $\pi^*$ has been followed up to this stage $t$. In fact, $\pi^*(\tilde{\theta}^t)$ might not even be optimal if $\tilde{\theta}^t \notin \tilde{\Theta}^t_{\pi^*}$. This in turn means that $Q^*(\tilde{\theta}^{t-1},a)$, the Q-values for predecessors of $\tilde{\theta}^t$, might not be the optimal.

### 3.1.5 Optimal Dec-POMDP Value Functions

We demonstrated that the optimal Q-value function for a Dec-POMDP is not well-defined without assuming a past joint policy. We propose a definition of the optimal value function of a Dec-POMDP that explicitly incorporates the past joint policy.

**Theorem 3.2 (Optimal $Q^*$).** The optimal Q-value function is properly defined as a function of the initial state distribution and joint past policies, action-observation histories and decision rules $Q^*(b^0,\varphi^t,\tilde{\theta}^t,\delta^t)$. This $Q^*$ specifies the optimal value given for all $\tilde{\theta}^t$, even for $\tilde{\theta}^t$ that are not reached by execution of an optimal joint policy $\pi^*$.

**Proof.** For all $\tilde{\theta}^t,\varphi^t,\delta^t$, the optimal expected return, respectively for the last stage and for all $0 \leq t < h-1$, is given by

$$Q^*(b^0,\varphi^{h-1},\tilde{\theta}^{h-1},\delta^{h-1}) = R(\tilde{\theta}^{h-1},\delta^{h-1}(\tilde{\theta}^{h-1})), \quad (3.1.12)$$

$$Q^*(b^0,\varphi^t,\tilde{\theta}^t,\delta^t) = R(\tilde{\theta}^t,\delta^t(\tilde{\theta}^t)) + \sum_{o^t+1} \Pr(o^t+1|\theta^t,\delta^t(\theta^t))Q^*(b^0,\varphi^{t+1},\tilde{\theta}^{t+1},\delta^{t+1,*}), \quad (3.1.13)$$

\(^1\)The question as to how the mistake of one agent should be detected by another agent is a different matter altogether and beyond the scope of this text.
with $\varphi^{t+1} = (\varphi^t \circ \delta^t)$ and

$$
\delta^{t+1,*} = \arg \max_{\delta^{t+1}} \sum_{\tilde{\theta}^{t+1} \in \Theta^{t+1}} \Pr(\tilde{\theta}^{t+1}|b^0, \varphi^{t+1}) Q^*(b^0, \varphi^{t+1}, \tilde{\theta}^{t+1}, \delta^{t+1}).
$$

These definitions are consistent. Because of (3.1.12) for the last stage (3.1.14) will maximize the expected reward and thus is optimal. Equation (3.1.13) propagates these optimal value to the preceding stage. As such optimality for all stages follows by induction.

The above equations constitute a dynamic program. When assuming that only pure joint past policies $\varphi$ can be used, the dynamic program can be evaluated from the end ($t = h - 1$) to the beginning ($t = 0$). Figure 3.3 illustrates the computation of $Q^*$.

When arriving at stage 0, the past joint policy is empty $\varphi^0 = ()$ and joint decision rules are simply joint actions, thus it is possible to select

$$
\delta^{0,*} = \arg \max_{\delta^0} Q^*(b^0, \varphi^0, \tilde{\theta}_0, \delta^0) = \arg \max_a Q^*(b^0, (), \tilde{\theta}_0, a).
$$

Then given $\varphi^1 = \delta^{0,*}$ we can determine $\delta^{1,*}$ using (3.1.14), etc. This essentially is the forward-sweep policy computation using optimal Q-value function as defined by (3.1.13). Note that the solution of the Bayesian game (i.e., performing the maximization in (3.1.14)) has already been done and can be cached.

At this point, there is no clear understanding of $Q^*(b^0, \varphi^t, \tilde{\theta}_t, \delta_t)$ as an action-value function anymore, i.e., it is no longer coupled to domain level (joint) actions. Of course it is possible to recover these. The value of performing $a = \delta^t(\tilde{\theta}^t)$ from
\[ \tilde{\theta}^t \] after having followed \( \varphi^{t+1} = \varphi^t, \delta^t \) and continuing optimally afterward is given by
\[ Q_{\varphi^{t+1}}(\tilde{\theta}^t, \delta^t) = Q^*(b^0, \varphi^t, \tilde{\theta}^t, \delta^t). \] (3.1.16)

This can be used for instance to compute \( Q_{\pi^*} \).

Another emerging question is why we still refer to a Q-value function, i.e., why we use the symbol 'Q' to refer to the value function defined here. The answer is that \( \delta^t \) can be seen as an action on the meta-level of the planning process. In the same way we can interpret \( \varphi^t \) as the state in this planning process and we can define \( V \) and \( Q \) with their usual interpretations. In particular, it is possible to write
\[ V^*(b^0, \varphi^t) = \max_{\delta^t} Q^*(b^0, \varphi^t, \delta^t) \] (3.1.17)

where \( Q^* \) is defined as
\[ Q^*(b^0, \varphi^t, \delta^t) = \sum_{\tilde{\theta}^t} \Pr(\tilde{\theta}^t|b^0, \varphi^t)Q^*(b^0, \varphi^t, \tilde{\theta}^t, \delta^t). \] (3.1.18)

Note that this \( Q^* \) indeed has the regular interpretation of the expected immediate reward induced by first taking 'action' \( \delta^t \) plus the cumulative reward of continuing optimally afterward. We can see this by rewriting
\[
\begin{align*}
Q^*(b^0, \varphi^t, \delta^t) &= \sum_{\tilde{\theta}^t} \Pr(\tilde{\theta}^t|b^0, \varphi^t) \\
&= \sum_{\tilde{\theta}^t} \Pr(\tilde{\theta}^t|b^0, \varphi^t) \left[ R(\tilde{\theta}^t, \delta^t(\tilde{\theta}^t)) + \sum_{o^{t+1}} \Pr(o^{t+1}|\tilde{\theta}^t, \delta^t(\tilde{\theta}^t))Q^*(b^0, (\varphi^t \circ \delta^t), \tilde{\theta}^{t+1}, \delta^{t+1,*}) \right] \\
&= \sum_{\tilde{\theta}^t} \Pr(\tilde{\theta}^t|b^0, \varphi^t)R(\tilde{\theta}^t, \delta^t(\tilde{\theta}^t)) + \sum_{\tilde{\theta}^t} \Pr(\tilde{\theta}^t|b^0, \varphi^t) \sum_{o^{t+1}} \Pr(o^{t+1}|\tilde{\theta}^t, \delta^t(\tilde{\theta}^t))Q^*(b^0, (\varphi^t \circ \delta^t), \tilde{\theta}^{t+1}, \delta^{t+1,*}) \\
&= E[R(s^t, a^t) | b^0, \varphi^t, \delta^t] + \sum_{\tilde{\theta}^{t+1}} \Pr(\tilde{\theta}^{t+1}|b^0, (\varphi^t \circ \delta^t))Q^*(b^0, (\varphi^t \circ \delta^t), \tilde{\theta}^{t+1}, \delta^{t+1,*}) \\
&= E[R(s^t, a^t) | b^0, \varphi^t, \delta^t] + Q^*(b^0, (\varphi^t \circ \delta^t), \delta^{t+1,*}) \\
&= E[R(s^t, a^t) | b^0, \varphi^t, \delta^t] + \max_{\delta^{t+1}} Q^*(b^0, (\varphi^t \circ \delta^t), \delta^{t+1}) \\
&= E[R(s^t, a^t) | b^0, \varphi^t, \delta^t] + V^*(b^0, (\varphi^t \circ \delta^t)). \quad (3.1.19)
\end{align*}
\]

Here \( E[R(s^t, a^t) | b^0, \varphi^t, \delta^t] \) corresponds to the expected immediate reward at stage \( t \) and \( V^*(b^0, (\varphi^t \circ \delta^t)) \) to the optimal value of continuing afterward.

### 3.1.5.1 The Relation to (Point-Based) Dynamic Programming

The computation of \( Q^* \) is closely related to (point-based) dynamic programming for Dec-POMDPs as discussed in Subsection 2.6.5 and 2.6.5.1. Suppose that \( t = 2 \).
in Figure 3.3 is the last stage (i.e., $h = 3$). When for each $\varphi^2 = \langle \varphi^1 \circ \delta^1 \rangle$ the maximizing $\delta^{2,*} = \langle \delta^{2,*}_1, \ldots, \delta^{2,*}_n \rangle$ has been computed, it is easy to construct the sets of non-dominated actions for each agent: every action $a_i$ of agent $i$ that is specified by some $\delta^{2,*}_i$ is non-dominated.

Once we have computed the values for all $\langle \theta^1, \varphi^2 = \langle \varphi^1 \circ \delta^1 \rangle \rangle$ at $t = 1$ (i.e., all $Q^*(b^0, \varphi^1, \theta^1, \delta^1)$ are computed), each $\varphi^2$ has an associated next-stage joint decision rule $\delta^{2,*}_2$, and in general an optimal joint future policy $\psi^1 = (\delta^{2,*}_2, \ldots, \delta^{h-1,*}_n)$.

This means that, given $\varphi^1, \delta^1$, we can define a joint sub-tree policy $q^{\tau=2}$ for each $\theta^1$. This is done by taking $\langle \delta^1 \circ \psi^1 \rangle$ and restricting this joint policy to the histories consistent with $\theta^1$. For instance, in Figure 3.3 the shaded trees represent the joint sub-tree policy for $\theta^1$ given the indicated past policy $\varphi^2 = \langle \varphi^1 \circ \delta^1 \rangle$. Clearly, $Q^*(b^0, \varphi^1, \theta^1, \delta^1)$ corresponds to the expected value of this associated joint sub-tree policy. Dynamic programming keeps track of these sub-trees policies. In contrast, the algorithm to compute $Q^*$ presented in this section keeps track of the values.

### 3.1.5.2 The Relation to Sequential Rationality

The fact that it is not possible to simply define $Q^*(\bar{\theta}^t, a^t)$ is very much related to the notion of sub-game perfect equilibria from game theory. As explained in Section (2.2), a sub-game perfect Nash equilibrium $\pi = \langle \pi_1, \ldots, \pi_n \rangle$ has the characteristic that the contained policies $\pi_i$ specify an optimal action for all possible situations—even situations that can not occur when following $\pi$. A commonly given rationale behind this concept is that, by a mistake of one of the agents during execution, situations that should not occur according to $\pi$, may occur, and also in these situations the agents should act optimally. A different rationale is given by Binmore (1992), who remarks that although it is tempting to ignore situations that should not occur according to $\pi$, it would clearly be a mistake, because the agents “remain on the equilibrium path because of what they anticipate would happen if they were to deviate”. This implies that agents can decide upon a Nash equilibrium by analyzing what the expected outcome would be by following other policies: that is, when acting optimally from other situations. In the previous section, we performed a similar reasoning for Dec-POMDPs, which—in a similar fashion—resulted in a description that allows to deduce an optimal Q-value function and thus joint policy.

A Dec-POMDP can be modeled as an extensive form game of imperfect information (Oliehoek and Vlassis, 2006). For such games, the notion of sub-game perfect equilibria is inadequate; because this type of games often do not contain proper sub-games, every Nash equilibrium is trivially sub-game perfect. (The extensive form of a Dec-POMDP indeed does not contain proper sub-games, because agents can never discriminate between the other agents’ observations.) To overcome this problem different refinements of the Nash equilibrium concept have been defined, of which we will mention the assessment equilibrium (Binmore, 1992) and the closely

---

1In this example $\psi^1 = \delta^{2,*}$ because $h = 3$

2Remember a $\tau$-stages-to-go sub-tree policy is rooted at stage $t = h - \tau$, so $q^{\tau=2}$ starts at stage $t = 3 - 2 = 1$. 

---
related, but stronger *sequential equilibrium* (Osborne and Rubinstein, 1994). Both
these equilibria are based on the concept of an assessment, which is a pair \((\pi, B)\)
consisting of a joint policy \(\pi\) and a *belief system* \(B\). The belief system maps the
information sets of each agent—also the ones that are not reachable given \(\pi\)—to a
probability distributions over nodes and thus possible joint histories. Roughly
speaking, an assessment equilibrium requires *sequential rationality* and *belief con-
sistency*.\(^1\) The former entails that the joint policy \(\pi\) specifies optimal actions for
each information set given \(B\). Belief consistency means that all the beliefs that are
assigned by \(B\) are Bayes-rational given the specified joint policy \(\pi\), i.e., the beliefs
are computed through proper application of Bayes’ rule.\(^2\)

This is in direct correspondence to (3.1.14), which specifies to follow a strategy
that is rational given the probabilities of all possible histories \(\tilde{\theta}\), and those probabil-
ities (the corresponding consistent belief system) are computed correctly. As such
we also refer to \(Q^*\) as defined in Theorem 3.2 as the ‘sequentially rational’ Q-value
function, in contrast to the normative description \(Q^*_\pi\) of (3.1.6). Also note that
using \(Q^*\) the optimal future policy can be computed for *any* past policy. This may
have important applications in an online setting. For instance, suppose agent \(i\)
makes a mistake at stage \(t\), executing an action not prescribed by \(\pi^*_i\), assuming the
other agents execute their policy \(\pi_{\neq i}\) without mistakes, agent \(i\) knows the actually
executed previous policy \(\varphi^{t+1}\). Therefore it can compute a new individual policy by

\[
\delta^{t+1}_i = \arg \max_{\delta^{t+1}} \sum_{\tilde{b}^{t+1}} \Pr(\tilde{b}^{t+1}|b^0, \varphi^{t+1}) Q^*(\tilde{b}^{t+1}, \varphi^{t+1}, (\delta^{t+1}_i, \delta^{t+1}_{\neq i})).
\]

### 3.1.5.3 The Complexity of Computing \(Q^*\)

Although we have now found a way to compute the optimal Q-value function, this
computation is intractable for all but the smallest problems. The last stage is
trivial, since the Q-values are given directly by the immediate reward function.
However, for stage \(t = h - 2\) the Q-value function \(Q^*(b^0, \varphi^t, \tilde{b}^t, \delta^t)\) has a huge
number of entries: effectively we need to compute an entry for each combination of
\(\varphi^{h-1} = (\varphi^{h-2} \circ \delta^{h-2})\) and each consistent joint action-observation history \(\tilde{\theta}^{h-2}\).

Up to and including an arbitrary stage \(t - 1\), there are \(\sum_{t'=0}^{t-1} |O_i|^{t'} = \frac{|O_i|^{t-1}}{|O_i|-1}\)
observation histories for agent \(i\) and thus the number of \(\varphi^t\) is

\[
O \left( |A_*|^{n(|O_*|^{t-1})} \right).
\]

For each of these past joint policies \(\varphi^t\) there are \(|\tilde{O}^t| = |O|^t\) consistent joint action-
observation histories at stage \(t\) (for each observation history \(\tilde{\theta}^t\), \(\varphi^t\) specifies the
actions forming \(\tilde{\theta}^t\)). This means that for stage \(h - 2\) (for \(h - 1\), the Q-values are

\(^1\)Osborne and Rubinstein (1994) refer to this second requirement as simply ‘consistency’. In
order to avoid any confusion with definition 3.1 we will use the term ‘belief consistency’.

\(^2\)A sequential equilibrium includes a more technical part in the definition of belief consistency
that addresses what beliefs should be held for information sets that are not reached according to
\(\pi\). For more information we refer to Osborne and Rubinstein (1994).
easily calculated), the number of entries to be computed is the number of joint past policies $\varphi^{h-1}$ times the number of joint histories
\[ O \left( |A|^n |O^{h-1}| \right), \]
indicating that computation of this function is doubly exponential in the horizon.
Also, for each joint past policy $\varphi^{h-2}, \delta^{h-2}$, we need to compute $\delta^{h-1,*}$ by solving a BG for the last stage. To the author’s knowledge, the only method to optimally solve these BGs is evaluation of all $O(\prod |A^*|)$ joint BG-policies.

As such, computing optimal value functions is intractable. In the remainder of this chapter we will treat value functions for Dec-POMDPs with communication and show that these are easier to compute. In the next chapter, we will consider using approximate value functions for the non-communicative setting.

### 3.2 Instantaneous Communication

This section describes the optimal value functions for the setting in which the agents are capable of instantaneous, noise-free and cost-free communication. In particular Pynadath and Tambe (2002b) showed that under such circumstances, it is optimal for the agents to share their local observations (e.g., by broadcasting the individual observations to all other agents).

In effect such communication transforms a Dec-POMDP to what we refer to as a **multiagent POMDP (MPOMDP)**.\(^1\) We can think of the MPOMDP as a POMDP in which there is a puppeteer agent that takes joint actions and receives joint observations. By solving this underlying POMDP of the Dec-POMDP, all POMDP-literature applies to this setting.

In particular, the optimal 0-steps delay value function $V_0$ for an underlying POMDP satisfies:
\[ V_0^* (b') = \max_a Q_0^*(b',a) \]
\[ Q_0^*(b',a) = R(b',a) + \sum_{o^{t+1} \in \mathcal{O}} P(o^{t+1} | b', a) \max_{a^{t+1}} Q_0^*(b^{t+1}, a^{t+1}), \]
where $b'$ is the joint belief (corresponding to some joint AOH $\tilde{\theta}^t$) of the single puppeteer agent that selects joint actions and receives joint observations at time step $t$, where
\[ R(b',a) = \sum_{s \in \mathcal{S}} R(s,a) b^t(s) \]
is the expected immediate reward, and where $b^{t+1}$ is the joint belief resulting from $b'$ by action $a$ and joint observation $o^{t+1}$, calculated by Bayes’ rule:
\[ \forall s', b^{t+1}(s') = \frac{\Pr(o | a, s') \sum_{s \in \mathcal{S}} \Pr(s | s, a) b^t(s)}{\sum_{s' \in \mathcal{S}} \Pr(o | a, s') \sum_{s \in \mathcal{S}} \Pr(s' | s, a) b^t(s)}. \]

\(^1\)This name is chosen in analogy with the multiagent MDP, described in Subsection 2.8.1.
3.2 Instantaneous Communication

For a finite horizon, $Q_0^*$ can be computed by generating all possible joint beliefs and solving the belief MDP. Generating all possible beliefs is easy: starting with $b_0$ corresponding to the empty joint action-observation history $\bar{\theta}^{t=0}$, for each $a$ and $o$ we calculate the resulting $\theta^{t=1}$ and corresponding joint belief and continue recursively. Solving the belief MDP amounts to recursively applying (3.2.2).

The cost of computing the optimal MPOMDP Q-value function can be divided in the cost of calculating the expected immediate reward for all $\theta^{t,a}$, and the cost of evaluating future reward for all $\bar{\theta}^{t,a}$, with $t = 0, ..., h − 2$. The former operation has cost $O(|S|)$ per $(\theta^{t,a})$-pair. The latter requires selecting the maximizing joint action for each joint observation which induces a cost of $(|A||O|)$ per $(\bar{\theta}^{t,a})$-pair. The total complexity of computing $Q_0^*$ becomes

$$O \left( \frac{(|A||O|)^{h-1}}{(|A||O|) - 1} |A| (|A||O|) + \frac{(|A||O|)^{h-1}}{(|A||O|) - 1} |A||S| \right).$$

Equation (3.2.5)

Evaluating (3.2.2) for (joint beliefs for) all joint action-observation histories $\bar{\theta}^t \in \bar{\Theta}^t$ can be done in a single backward sweep through time. This can also be visualized in Bayesian games as illustrated in Figure 3.4; the expected future reward is calculated as a maximizing weighted sum of the entries of the next time step BG.\(^1\)

Nevertheless, solving a POMDP optimally is also known as an intractable problem. As a result, POMDP research in the last decade has focused on approximate solutions for POMDPs. In particular, it is known that the value function of a POMDP is piecewise-linear and convex (PWLC) over the (joint) belief space (Sondik, 1971). This property is exploited by many approximate POMDP solution methods (Pineau, Gordon, and Thrun, 2003; Spaan and Vlassis, 2005).

---

\(^1\)Note that the maximization in (3.2.2) can be seen as a special instance of the ‘BG-operator’ expressed by (3.1.14). Because, under instantaneous communication, it will be possible to prune all histories that are not realized from the BG, the BG-operator reduces to a simple maximization.
### 3.3 One-Step Delayed Communication

Here we describe the optimal value functions for a Dec-POMDP with noise-free and cost-free communication that arrives with a one-step delay (1-SD). I.e., the assumption is that during execution at stage $t$ the agents know $\vec{\theta}^{t-1}$, the joint action-observation history up to time step $t-1$, and the joint action $a^{t-1}$ that was taken at the previous time step. Because all the agents know $\vec{\theta}^{t-1}$, they can compute the joint belief $b^{t-1}$ it induces which is a Markovian signal. Therefore the agents do not need to maintain prior information, $b^{t-1}$ takes the same role as $b^0$ in a regular Dec-POMDP (i.e., without communication).

The 1-SD-setting has also been considered in the field of decentralized control, where it is usually referred to as decentralized control with a “one-step delayed sharing pattern”. In particular, Varaiya and Walrand (1978) showed that in this setting state estimation and control are separable. I.e., there exist an optimal separable joint policy, that specifies a separable individual policy for each agent that maps from $b^{t-1}$ and the individual history of observations received since $b^{t-1}$, to actions. Hsu and Marcus (1982) extended this work by deriving dynamic programming algorithms for the finite- and infinite-horizon case. Their approach, however, is based on the one-step predictor of $s^t$, leading to a rather involved formulation. In this section we present what we believe to be a conceptually clearer formulation, by resorting to Bayesian games.

As mentioned, in the 1-SD setting the agents know $\vec{\theta}^{t-1}$ and thus can compute $b^{t-1}$. Also, since we assume that during execution each agent knows the joint policy, each agent can defer the taken joint action $a^{t-1}$. However, the agents are uncertain regarding each other’s last observation, and thus regarding the joint observation $o^t$. Effectively, this situation defines a BG for each possible joint belief $b^{t-1}$ (induced by all possible $\vec{\theta}^{t-1}$) and joint action $a^{t-1}$. Note, however, that these BGs are different from the BGs used in Section 3.1.1: the BGs here have types that correspond to single observations, whereas the BGs in 3.1.1 have types that correspond to complete action-observation histories. Hence, the BGs of here are much smaller in size and thus easier to solve.

**Lemma 3.1** (Value of one-step delayed communication). The optimal value function for a finite-horizon Dec-POMDP with one-step delayed communication is given by

$$V_1^{t,*}(b^{t-1},a^{t-1}) = \max_{\beta^t} Q_1^{t,*}(b^{t-1},a^{t-1},\beta^t)$$

$$Q_1^{t,*}(b^{t-1},a^{t-1},\beta^t) = \sum_{o^t} \Pr(o^t|b^{t-1},a^{t-1})Q_1^{t,*}(b^{t-1},a^{t-1},o^t,\beta^t)$$

$$Q_1^{t,*}(b^{t-1},a^{t-1},o^t,\beta^t) = R(b^t,\beta^t(o^t)) + \max_{\beta^{t+1}} \sum_{o^{t+1}} \Pr(o^{t+1}|b^t,\beta^t(o^t))Q_1^{t+1,*}(b^t,\beta^t(o^t),o^{t+1},\beta^{t+1})$$

where $b^t$ results from $b^{t-1},a^{t-1},o^t$. 
Proof. \( Q_1^t(b^{t-1}, a^{t-1}, \beta^t) \) should be defined as the sum of the expected immediate and future reward, so (3.3.2) should equal

\[
E[R(s^t, a^t) | b^{t-1}, a^{t-1}, \beta^t] + E[V_{1}^{t+1,*}(b^t, a^t) | b^{t-1}, a^{t-1}, \beta^t] = \\
\sum_{o^t} \Pr(o^t | b^{t-1}, a^{t-1}) R(b^t, \beta^t(o^t)) + \sum_{o^t} \Pr(o^t | b^{t-1}, a^{t-1}) V_{1}^{t+1,*}(b^t, \beta^t(\tilde{o}^t)) \tag{3.3.4}
\]

Substitution of \( V_{1}^{t+1,*} \) yields

\[
\sum_{o^t} \Pr(o^t | b^{t-1}, a^{t-1}) \left[ R(b^t, \beta^t(o^t)) + \max_{\beta^{t+1}} \sum_{o^{t+1}} \Pr(o^{t+1} | b^t, \beta^t(o^t)) Q_{1}^{t+1,*}(b^t, \beta^t(o^t), o^{t+1}, \beta^{t+1}) \right] \tag{3.3.5}
\]

which shows that the definitions are consistent. For the last stage \( h - 1 \), the maximization in (3.3.1) assures that the future reward is maximized, as such (3.3.1) is optimal for the last stage. Given the optimality of the last stage, optimality for previous stages follows by induction: given the optimality of \( V_{1}^{h-1,*}, V_{1}^{h-2,*} \) maximizes the sum of the expected immediate reward at \( h - 2 \) plus the future reward \( V_{1}^{h-1,*} \), etc.

There is a clear correspondence between the equations in Lemma 3.1 and those in Section 3.1.5. In particular, (3.3.1) corresponds to (3.1.17), because \( a^{t-1} \) (resp. \( \varphi^t \)) is the joint policy taken since the last known Markovian signal (distribution over states) \( b^{t-1} \) (resp. \( b^0 \)). Similarly, (3.3.2) corresponds to (3.1.18) where \( \beta^t \) (resp. \( \delta^t \)) is the joint policy for stage \( t \) that implicitly maps sequences of joint observations \( o^t \) (resp. \( \tilde{o}^t \)) received since the last Markov signal to actions \( a^t \).

### 3.3.1 Immediate Reward Formulation

Note that the arguments of (3.3.3) are somewhat redundant. In particular, the right side only depends on the joint belief \( b^t \) that results from \( b^{t-1}, a^{t-1}, o^t \) and on the joint action \( a^t = \beta^t(\tilde{o}^t) \). As such it can be rewritten simpler as

\[
V_{1}^{t,*}(b^t, a^t) = R(b^t, a^t) + \max_{\beta^{t+1}} \sum_{o^{t+1} \in \mathcal{O}} \Pr(o^{t+1} | b^t, a^t) V_{1}^{t+1,*}(b^t, \beta^{t+1}(o^{t+1})) \tag{3.3.6}
\]

In doing so, we have now constructed a formulation that specifies the value over stages \( t, \ldots, h - 1 \) using arguments of stage \( t \). We refer to this formulation as an immediate reward value function. This should be seen in contrast to formulas of the form of (3.3.1) that specify the expected value over stages \( t, \ldots, h - 1 \), but using arguments (in particular a Markovian signal \( b^{t-1} \)) of stage \( t - 1 \) (or in general \( t - k \)). We refer to this type as expected reward value function, since they specify the expected value over later stages given arguments of an earlier stage.

The remainder of this chapter will use the expected reward formulation, but occasionally we will make a remark about the other immediate reward formulation,
as done here. We also note that in (3.3.6) we use \( V_1 \) and not \( Q_1 \) because this is consistent with the notation for immediate reward value function formulations. An overview of immediate reward value functions and the relation to expected reward formulations is given in Appendix B.

3.3.2 Complexity

The cost of computing \( V_{t,*}^{t,*}(b^t,a^t) \) for all \( \hat{\theta}^t, a \) can be split up in the cost of computing the immediate reward and the cost of computing the future reward (solving a BG over the last received observation), which is \( O(|A^*|^nO^*) \), leading to a total cost of:

\[
O\left(\left(\frac{|A||O|}{(|A||O|) - 1}\right)^{h-1} \cdot |A^*|^nO^* + \left(\frac{|A||O|}{(|A||O|) - 1}\right)^h - 1 |A|||S|\right)\). \tag{3.3.7}
\]

Comparing to the cost of computing \( V_0 \) given by (3.2.5), this contains an additional exponential term, but this term does not depend on the horizon of the problem.

As discussed in Section 3.2, \( V_0 \), the value function of the underlying POMDP, can be efficiently approximated by exploiting the PWLC-property of the value function. Hsu and Marcus (1982) showed that the value function of their formulation for 1-step delayed communication also preserves the PWLC property. Not surprisingly, \( V_{t,*}^{t,*}(b^t,a^t) \) in (3.3.6) is also PWLC over the joint belief space (Oliehoek et al., 2007c) and, as a result, approximation methods for POMDPs can be transferred to its computation (Oliehoek et al., 2007b).

3.4 \( k \)-Steps Delayed Communication

This section describes the setting where there is cost-free and noise-free communication, but there is a delay of \( k \) stages. Aicardi, Davoli, and Minciardi (1987) and Ooi and Wornell (1996) performed similar work on decentralized control in which there is a \( k \)-steps delayed state observation. That is, they consider the setting where, at time step \( t \), all agents \( i \) know their own observations \( o^0_i, \ldots, o^t_i \) and the states that have occurred up to \( t-k \): \( s^0, \ldots, s^{t-k} \). In particular Aicardi et al. (1987) consider the Dec-MDP setting in which agent \( i \)'s observations are local states \( \hat{s}_i \) and where a joint observation identifies the state \( s = (\hat{s}_1, \ldots, \hat{s}_n) \). As a consequence the delayed observation of the state \( s^{t-k} \) can be interpreted as the result of \( k \)-steps delayed communication in a Dec-MDP. Aicardi et al. present a dynamic programming formulation to optimally solve the finite-horizon problem. Ooi and Wornell (1996) examine the decentralized control of a broadcast channel over an infinite horizon and present a reformulation of the stochastic control problem, that is very close to the description that will be presented in this section. Ooi and Wornell extend upon the previous work by splitting the (individual) observations in a local and global part, lowering the complexity of evaluating the resulting dynamic program.
In this section we will present a reformulation and minor extension of previous work by demonstrating the solution for Dec-POMDPs where there is a k-steps delayed communication of the individual observation. That is, for systems where not the state, but the joint action-observation history is perceived with a k-stage delay:

\[ \text{at stage } t, \text{ each agent } i \text{ knows } \vec{\theta}_{t-k} \text{ and its individual } \theta_t^i. \]

For these systems the optimal value functions are discussed, reformulated to naturally fit in the overview of different communication assumptions. A main contribution is the result that decreased communication delays cannot decrease the expected value of decentralized systems in Section 3.4.4. For the centralized setting a similar result was shown by Bander and White (1999). Although it is a very intuitive result, and Ooi and Wornell (1996) use this intuition to motivate their approach, for the decentralized setting a formal proof had been lacking.

### 3.4.1 Modeling Systems with k-Steps Delay

In the setting of k-steps delayed communication, at stage \( t \) each agent agent knows \( \vec{\theta}_{t-k} \), the joint action-observation history of \( k \) stages earlier, and therefore can compute \( b_{t-k} \) the joint belief induced by \( \vec{\theta}_{t-k} \). Again, \( b_{t-k} \) is a Markov signal, so no further history needs to be retained and \( b_{t-k} \) takes the role of \( b_0 \) in the no-communication setting and \( b_{t-1} \) in the one-step delay setting. Indeed, one-step delay is just a special case of the \( k \)-steps delay setting.

In contrast to the one-step delayed communication case, the agents do not know the last taken joint action. However, since we assume the agents know each other policies, they do know \( q_{\tau=k,t-k} \), the joint policy that has been executed during stages \( t-k, \ldots, t-1 \). This \( q_{\tau=k,t-k} \) is a length-\( k \) joint sub-tree policy rooted at stage \( t-k \): it specifies a sub-tree policy \( q_{\tau=k,t-k}^i \) for each agent \( i \) which is a policy tree that specifies actions for \( k \) stages \( t-k, \ldots, t-1 \). \(^1\)

In this section we more concisely write \( q_{t-k}^{|k|} \) for a length-\( k \) joint sub-tree policy. The technicalities of how these sub-tree policies should be maintained are somewhat involved and require some more notation that will first be introduced. Afterward we will define optimal value functions and discuss the complexity of computing them.

Let us assume that at a particular stage \( t \) the situation is as depicted in the top half of Figure 3.5: the system with 2 agents has \( k = 2 \) steps delayed communication, so each agent knows \( b_{t-k} \) and \( q_{t-k}^{|k|} \) the joint sub-tree policy that has been executed during stages \( t-k, t-1 \). At this point the agents need to select an action, but they don’t know each others individual observation history since stage \( t-k \). That is they have uncertainty with respect to the length-\( k \) observation history \( \vec{\sigma}_{|k|} = (o_{t-k+1}, \ldots, o_t) \). Effectively, this means that the agents have to use a joint BG-policy \( \beta_{|k|} = (\beta_{1|k|}^t, \ldots, \beta_{n|k|}^t) \) that implicitly maps length-\( k \) observation histories to joint actions \( \beta_{|k|}(\vec{\sigma}_{|k|}) = a^t \).

We assume that in the planning phase we computed such a joint BG-policy \( \beta_{|k|}^t \)

---

\(^1\)Remember from Subsection 2.6.5 that we use \( \tau \) to denote the number of steps-to-go and in the context of a sub-tree policy \( \tau \) refers to its ‘length’: the number of time-steps for which it specifies actions.
as indicated in the figure. As is shown, \( \beta_{t|k}^i \) can be used to extend the sub-tree policy \( q_{t|k}^{t-k} \) to form a longer sub-tree policy with \( \tau = k + 1 \) stages-to-go. Each agent has knowledge of this extended joint sub-tree policy

\[
q_{t|k+1}^{t-k} = (q_{t|k}^{t-k} \circ \beta_{t|k}^i).
\]

Consequently each agent \( i \) executes the action corresponding to its individual observation history \( \beta_{t|k}^i(\alpha_{t|k}^i) = a_t^i \) and a transition occurs to stage \( t + 1 \). At that point each agent receives a new observation \( o_{t+1}^i \) through perception and the joint observation \( o_{t-k+1}^t \) through communication, it transmits its individual observation, and computes \( b_{t-k+1}^i \). Now, all the agents know what action was taken at \( t - k \) and what the following observation \( o_{t-k+1}^t \) was. Therefore the agents know which part of \( q_{t|k+1}^{t-k} \) has been executed during the last \( k \) stages \( t - k + 1, \ldots, t \) and they discard the part not needed further. I.e., the joint observation ‘consumes’ part of the joint sub-tree policy.

**Definition 3.2 (Policy consumption).** Feeding a length-\( k \) joint sub-tree policy \( q \) with a sequence \( l < k \) joint observations consumes a part of \( q \) leading to a joint sub-tree policy \( q' \) which is a sub-tree of \( q \). In particular, consumption \( \Downarrow \) by a single joint observation \( o_{t-k+1}^t \) is written as

\[
q_{t|k+1}^{t-k} = q_{t|k+1}^{t-k} \Downarrow o_{t-k+1}^t. \tag{3.4.1}
\]
This process is illustrated in the bottom part of Figure 3.5. Policy consumption also applies to joint BG-policies:

\[ \beta^t_{|k-1|} = \beta^t_{|k|} \bigg\|_{o^{t-k+1}} \]

**Proposition 3.3** (Distributivity of policy operations). Policy concatenation and consumption are distributive. That is

\[ \langle q^{t-k}_{|k|} \bigg\|_{o^{t-k+1}} \circ \beta^t_{|k|} \bigg\|_{o^{t-k+1}} \rangle = \langle q^{t-k}_{|k|} \circ \beta^t_{|k|} \bigg\|_{o^{t-k+1}} \rangle \]

**Proof.** This statement can easily be verified by inspection of Figure 3.5. A formal proof is omitted.

### 3.4.2 Optimal Value Functions

This section discusses the optimal value functions under \( k \)-steps delayed communication. To ease notation we will simply write \( q^{t-k}, \beta^t \) for \( q^{t-k}_{|k|}, \beta^t_{|k|} \). With some abuse of notation we write \( \beta^t(\bar{\theta}^t_{|k|}) \) to denote \( \beta^t_{|k|}(o^{t-k+1}, \ldots, o^t) \) the application of the length-\( k \) joint BG policy to the last \( k \) joint observations of \( \bar{\theta}^t_{|k|} \).

Also we will consider probabilities of the form \( \Pr(\bar{\theta}^t_{|k|} | b^{t-k}, q^{t-k}) \). These are defined as marginals of \( \Pr(s^t, \bar{\theta}^t_{|k|} | b^{t-k}, q^{t-k}) \) that are defined analogous to (2.5.6)

\[
\Pr(s^t, \bar{\theta}^t_{|k|} | b^{t-k}, q^{t-k}) = \Pr(o^t | a^{t-1}, s^t) \sum_{s^{t-1}} \Pr(s^t | s^{t-1}, a^{t-1}) \\
\quad \times \Pr(a^{t-1} | \bar{\theta}^{t-1}_{|k-1|}, q^{t-k}) \Pr(s^{t-1}, \bar{\theta}^{t-1}_{|k-1|} | b^{t-k}, q^{t-k})
\]

(3.4.2)

In a similar way as the no-communication and one-step delayed communication settings we have the following value function in the \( k \)-steps delayed communication case. That is, the ‘start point’ is \( b^{t-k} \) instead of \( b^0 \) or \( b^{t-1} \) and \( q^{t-k} \) takes the role of respectively \( \varphi^t \), \( a^{t-1} \).

**Lemma 3.2** (Value of \( k \)-steps delayed communication). The optimal value function for a finite-horizon Dec-POMDP with \( k \)-steps delayed, cost and noise free, communication is given by:

\[
V^{t,*}_{k} (b^{t-k}, q^{t-k}) = \max_{\beta^t} Q^{t,*}_{k} (b^{t-k}, q^{t-k}, \beta^t).
\]

(3.4.3)

\[
Q^{t,*}_{k} (b^{t-k}, q^{t-k}, \beta^t) = \sum_{\bar{\theta}^t_{|k|}} \Pr(\bar{\theta}^t_{|k|} | b^{t-k}, q^{t-k}) Q^{t,*}_{k} (b^{t-k}, q^{t-k}, \bar{\theta}^t_{|k|}, \beta^t)
\]

(3.4.4)

\[
Q^{t,*}_{k} (b^{t-k}, q^{t-k}, \bar{\theta}^t_{|k|}, \beta^t) = R(b^t, \beta^t(\bar{\theta}^t_{|k|})) + \\
\sum_{o^{t+1}} \Pr(o^{t+1} | b^t, \beta^t(\bar{\theta}^t_{|k|})) Q^{t+1,*}_{k} (b^{t-k+1}, q^{t-k+1}, \bar{\theta}^{t+1}_{|k|}, \beta^{t+1})
\]

(3.4.5)
where \( b' \) results from \( b^{t-k}, \tilde{\theta}^t_{|k|} \), and

\[
\beta^{t+1,*} = \arg \max_{\beta^{t+1}} \sum_{\tilde{\theta}^t_{|k|}} \Pr(\tilde{\theta}^{t+1}_{|k|} | b^{t-k+1}, q^{t-k+1} ) Q^{t+1,*}_{k} (b^{t-k+1}, q^{t-k+1}, \tilde{\theta}^{t+1}_{|k|}, \beta^{t+1})
\]  
(3.4.6)

**Sketch of proof.** We will show that

\[
Q^{t,*}_{k} (\tilde{\theta}^{t-k}, q^{t-k}, \beta^t) = E[R(s^t, a^t) | \tilde{\theta}^{t-k}, q^{t-k}, \beta^t] + E[V^{t+1,*}_{k} (\tilde{\theta}^{t-k+1}, q^{t-k+1}) | \tilde{\theta}^{t-k}, q^{t-k}, \beta^t]
\]  
(3.4.7)

and that the other equations are consistent with this definition. This means that, for the last stage (3.4.3) maximizes the expected reward and therefore is optimal, optimality for other stages follows immediately by induction. The full proof is listed in the Appendix D.

Again, there is a clear correspondence between the equations presented here and those for the no-communication and one-step delayed communication setting. In particular, (3.4.3) corresponds to (3.3.1), (3.1.17) and (3.4.4) corresponds to (3.3.2),(3.1.18). More relations between the value functions are discussed in Appendix B.

### 3.4.3 Complexity

The equations in Lemma 3.2 form a dynamic program that can be evaluated from end to begin. We give an analysis of the complexity by considering the number of entries of \( V^{h-1,*}_{k} (b^{h-1-k}, q^{h-1-k}) \) and the amount of work needed to compute one entry. The number of length-\( k \) joint sub-tree policies \( q \) is

\[
O \left( |A_*|^{n(|O_*|^{k-1})} \right) \tag{3.4.8}
\]

where \( |O_*|^{k-1} = \sum_{t=0}^{k-1} |O_*|^t \). The number of joint beliefs \( b \) is bounded by the number of \( \tilde{\theta} \) that induce those joint beliefs, is given by:

\[
\sum_{t=0}^{h-1-k} |\tilde{\theta}^t| = \sum_{t=0}^{h-1-k} (|A| |O|)^t = \frac{(|A| |O|)^{h-k}}{(|A| |O|) - 1}.
\]  
(3.4.9)

The maximization over \( \beta \) needs to consider \( |A_*|^{n|O_*|^{k}} \) joint policies. As a result, complexity induced by selecting the best \( \beta \) for each \( b^{h-1-k}, q^{h-1-k} \) is

\[
O \left( \frac{(|A| |O|)^{h-k}}{(|A| |O|) - 1} \cdot |A_*|^{n|O_*|^{k}} \cdot |A_*|^{n|O_*|^{k}} \right) \tag{3.4.10}
\]
which should be compared to the first term in (3.3.2). The total complexity is the sum of (3.4.10) and the complexity induced by computing the expected immediate rewards. Careful inspection of equations in Lemma 3.2 reveals that for each $b^t$ induced by a pair $(b^{t-k}, \theta^t_{[k]})$ the expected immediate reward will need to be computed for each joint action. The number of such $b^t$ is again bounded by the number of joint AOHs. Therefore, the complexity induced by these computations is

$$O\left(\frac{|A| |O|}{|A| - 1} - 1 |A| |S|\right)$$

as before in (the right side of) (3.3.2).

In the sections on immediate and 1-step delayed communication, we mentioned that because these value functions are PWLC over the joint belief space, we could use efficient approximate computation methods. Clearly it would be beneficial to also perform approximate computation for $V_k$. Unfortunately it turns out that this is not straightforward: for $k \geq 2$ Varaiya and Walrand (1978) showed that $k$-steps delayed communication systems\(^1\) are not separable. The result of this is that, just as in the Dec-POMDP case (Oliehoek et al., 2007c), $V_k$ is not a function of the joint belief space, let alone PWLC over this space.

### 3.4.4 Less Delay Cannot Decrease Value

This section shows that when the delay of communication decreases, the expected value cannot become less. That is, less delay in communication is not harmful. A related result by Bander and White (1999) shows that for a single agent POMDP with delayed observations, decreasing delays will not decrease the expected return. For the decentralized case, however, no such results were available.

**Theorem 3.3** (Shorter communication delays cannot decrease the expected value). The expected value over stages $t, \ldots, h - 1$ given a joint belief $b^{t-k-1}$ and joint policy $q^{t-k-1}_{[k+1]}$ followed during stages $t - k - 1, \ldots, t - 1$ is no less under $k$-steps communication delay, than under $(k+1)$-steps delay. That is

$$\forall t \forall b^{t-k-1} \forall q^{t-k-1}_{[k+1]} \ E[V^t_{k+1}(b^{t-k-1}, q^{t-k-1}_{[k+1]}) | b^{t-k}, q^{t-k}_{[k]} \geq V^t_k(b^{t-k-1}, q^{t-k-1}_{[k+1]}),$$

where the expectation on the left-hand side is over joint observations $o^{t-k}$ that together with $b^{t-k-1}, q^{t-k-1}_{[k+1]}$ induce $b^{t-k}$.

**Sketch of proof.** The proof is by induction. The base case is that (3.4.11) holds for the last stage $t = h - 1$. The full proof is listed in Appendix D.

### 3.5 Conclusions

A large body of work in single-agent decision-theoretic planning is based on value functions, but such theory has been lacking thus far for Dec-POMDPs. Given

---

\(^1\)Varaiya and Walrand refer to $k$ units delayed ‘sharing patterns’.
the large impact of value functions on single-agent planning under uncertainty, we expect that a thorough study of value functions for Dec-POMDPs can greatly benefit multiagent planning under certainty. This chapter presented a framework of Q-value functions for Dec-POMDPs under different communication assumptions, providing a significant contribution to fill this gap in Dec-POMDP theory.

The main contributions are for the setting without communication, where an optimal joint policy \( \pi^* \) induces an optimal Q-value function \( Q_{\pi^*}(\tilde{\theta}^t,a) \), and how it is possible to construct an optimal joint policy \( \pi^* \) using forward-sweep policy computation. This entails solving Bayesian games for time steps \( t = 0, \ldots, h - 1 \) which use \( Q_{\pi^*}(\tilde{\theta}^t,a) \) as the payoff function. Because \( Q_{\pi^*} \) implicitly depends on the optimal past joint policy, there is no clear way to compute \( Q_{\pi^*}(\tilde{\theta}^t,a) \) directly.

To overcome this problem, we introduced a different description of the optimal Q-value function \( Q^*(b_0,\varphi^{h-1},\tilde{\theta}^{h-1},\delta^{h-1}) \) this makes the dependence on the past joint policy explicit. This new description of \( Q^* \) can be computed using dynamic programming and can then be used to construct \( \pi^* \).

Another important contribution of this chapter is that it shows that a decrease in communication delay cannot lead to a decrease in expected return. That is, shorter communication delays are not harmful.

Finally, the chapter has presented a unified overview of the optimal value functions under various delays of communication and discussed how they relate to each other. Two formulations, the expected reward and immediate reward formulation, were identified, of which the former is used in this chapter. The latter form and the relations between the two is explained in Appendix B.

### 3.6 Future Work

An interesting direction of future work is to try to extend the results of this chapter to partially observable stochastic games (POSGs) (Hansen et al., 2004), which are Dec-POMDPs with an individual reward function for each agent. Since the dynamics of the POSG model are identical to those of a Dec-POMDP, a similar modeling via Bayesian games is possible when allowing the past policy to be stochastic. An interesting question is whether also in this case, an ‘optimal’ joint policy (i.e., a Pareto-optimal Bayes-Nash equilibrium) can be found by forward-sweep policy computation.