Value-Based Planning for Teams of Agents in Stochastic Partially Observable Environments

Oliehoek, F.A.

Citation for published version (APA):
Appendix B

Immediate Reward Value Function Formulations

In Chapter 3, we discussed that there are two ways of formulation value functions for decentralized settings: the expected reward and immediate reward formulations. The former were treated in Chapter 3. Here we treat the latter, and discuss the relation between the two formulations.

B.1 $k$-Steps Delay Immediate Reward Formulation

As in the 1-step delay case, it is possible to rephrase the value functions for $k$-steps delay to an immediate reward (IR) formulation, as described by Oliehoek et al. (2008b). Again, in this immediate reward formulation, the value $V_{t,*}^{k}$ over stages $t,\ldots,h-1$ is expressed in terms of arguments for stage $t$: $\forall_0\leq t\leq h-k-1$

$$Q_k^{t,*}(b^t,q^t,\beta^{t+k}) = R(b^t,q^t(\vec{o}_\emptyset)) + \sum_{o^{t+1}\in O} \Pr(o^{t+1}|b^t,q^t(\vec{o}_\emptyset))V_{t+1,*}^{k}(b^{t+1},q^{t+1}),$$

with $q^t(\vec{o}_\emptyset) = a^t$ the joint action specified for the empty joint history, $q^{t+1} = \langle q^t \circ \beta^{t+k} \rangle\|_{o^{t+1}}$ and

$$V_k^{t,*}(b^t,q^t) \equiv \max_{\beta^{t+k}} Q_k(b^t,q^t,\beta^{t+k}).$$

For the last $k$ stages, $h-k \leq t \leq h-1$, there are $\tau = h-t$ stages to go and we get

$$V_k^{t,*}(b^t,q^t_{|\tau}) = R(b^t,q^t_{|\tau}(\vec{o}_\emptyset)) + \sum_{o^{t+1}} \Pr(o^{t+1}|b^t,q^t_{|\tau}(\vec{o}_\emptyset))V_k^{t+1,*}(b^{t+1},q^{t+1}_{|\tau-1}).$$

Note that (B.1.3) does not include a maximization over ‘actions’ $\beta^{t+k}$. Therefore, the last $k$ stages should be interpreted as a Markov chain. Standard dynamic
programming can be applied to calculate all $V_{k}^{t,*}(b^{t}, q^{t})$-values, for all (joint beliefs induced by) all joint action-observation histories.

### B.2 Conversion between Formulations

Clearly the immediate and expected reward formulations should have some relation to each other. In particular, the IR-value $V_{k}^{t,*}(b^{t}, q^{t})$ can be decomposed into as the expected reward over the first $k$-steps $t, \ldots, t+k-1$, which we will denote using $K_{k}$, and the expected reward over the remaining steps, i.e., the expected reward (ER) formulation $V_{k}^{t+k,*}(b^{t}, q^{t})$:

$$V_{k}^{t,*}(b^{t}, q^{t}) = K_{k}(b^{t}, q^{t}) + V_{k}^{t+k,*}(b^{t}, q^{t}). \quad (B.2.1)$$

In the following, we define, $K_{k}(b^{t}, q^{t})$ through $K_{\tau}=i(\vec{\theta}_{t}, q_{\tau}=i, t)$, the expected reward for the next $i$st stages, i.e.,

$$K_{k}(b^{t}, q^{t}) \equiv K_{\tau}=k(b^{t}, q^{k}_{t}). \quad (B.2.2)$$

We then have $K_{\tau}=1(b^{t}, a^{t}) = R(b^{t}, a^{t})$ and

$$K_{\tau}=i(b^{t}, q^{t}_{i}) = R(b^{t}, q^{t}_{i}(\vec{\theta}_{0}))+ \sum_{o^{t+1}} \Pr(o^{t+1}|b^{t}, q^{t}_{i}(\vec{\theta}_{0}))K_{\tau}=i-1(b^{t+1}, q^{t+1}_{i-1}), \quad (B.2.3)$$

where $q^{t+1}_{i-1} = q^{t}_{i} \upharpoonright o^{t+1}$ is the depth-$(i-1)$ joint policy that results from $q^{t}_{i}$ after joint observation of $o^{t+1}$.

### B.3 Less Delay Cannot Decrease Value

Also for the IR formulation, the intuitive result that less delay cannot hurt, holds.

**Theorem B.1** (Shorter communication delays cannot decrease the value). The optimal Q-value function $Q_{k}$ of a finite horizon Dec-POMDP with $k$-steps delayed communication is an upper bound to $Q_{k+1}$, that of a $k+1$-steps delayed communication system. That is

$$\forall t \forall b^{t} \forall q^{t}_{k} \beta^{t+k}_{k} Q^{t,*}_{k}(b^{t}, q^{t}_{k} \beta^{t+k}_{k}) \geq \max_{\beta^{t+k}_{k+1}} Q^{t,*}_{k+1}(b^{t}, q^{t}_{k} \beta^{t+k}_{k+1}\beta^{t+k+1}_{k+1}). \quad (B.3.1)$$

**Proof.** We start be rewriting the left hand side:
\[ Q_{k+1}^{t,*}(b^t, q^t_k, \beta^{t+k}) \]
\[ = R(b^t, q^t_k, (\emptyset)) + \sum_{o^{t+1} \in \Omega} \Pr(o^{t+1}|b^t, q^t_k, (\emptyset))V_{k+1}^{t+1,*}(b^{t+1}, q^t_k, \beta^{t+k}) \]
\[ = R(b^t, q^t_k, (\emptyset)) + \sum_{o^{t+1} \in \Omega} \Pr(o^{t+1}|b^t, q^t_k, (\emptyset)) \]
\[ = \left[ K_k(b^{t+1}, q^t_k, \beta^{t+k}) + V_{k+1}^{t+1,*}(b^{t+1}, q^t_k, \beta^{t+k}) \right] \]
\[ = \left[ R(b^t, q^t_k, (\emptyset)) + \sum_{o^{t+1} \in \Omega} \Pr(o^{t+1}|b^t, q^t_k, (\emptyset))K_k(b^{t+1}, q^t_k, \beta^{t+k}) \right] + \]
\[ \sum_{o^{t+1} \in \Omega} \Pr(o^{t+1}|b^t, q^t_k, (\emptyset))V_{k+1}^{t+1,*}(b^{t+1}, q^t_k, \beta^{t+k}) \] (B.3.2)

Note that, for the right hand side, we have that
\[ \max_{\beta^{t+k}_{|k|}} Q_{k+1}^{t,*}(b^t, q^t_k, \beta^{t+k}) = V_{k+1}^{t,*}(b^t, q^t_k, \beta^{t+k}) \] (B.3.4)
and therefore we can write
\[ V_{k+1}^{t,*}(b^t, q^t_k, \beta^{t+k}) = K_{k+1}(b^t, q^t_k, \beta^{t+k}) + V_{k+1}^{t+k+1,*}(b^t, q^t_k, \beta^{t+k}) \] (B.3.5)
where, per definition (by (B.2.2) and (B.2.3))
\[ K_{k+1}(b^t, q^t_k, \beta^{t+k}) = K_{\tau=k+1}(b^t, q^t_k, \beta^{t+k}) = R(b^t, q^t_k, \beta^{t+k}, (\emptyset)) \]
\[ + \sum_{o^{t+1}} \Pr(o^{t+1}|b^t, q^t_k, \beta^{t+k}, (\emptyset))K_{\tau=k+1}(b^{t+1}, q^t_k, \beta^{t+k}) \] (B.3.6)

Note that in this equation \( \beta^{t+k}_{|k|} \) does not influence the immediate reward or observation probability. Therefore the last equation is equal to the first part of (B.3.3).
This means that we only have to show that
\[ \sum_{o^{t+1} \in \Omega} \Pr(o^{t+1}|b^t, q^t_k, (\emptyset))V_{k+1}^{t+k+1,*}(b^{t+1}, q^t_k, \beta^{t+k}) \geq V_{k+1}^{t+k+1,*}(b^t, q^t_k, \beta^{t+k}) \] (B.3.7)
However, this is exactly what Theorem 3.3 shows. Therefore (B.3.7) holds, concluding the proof.

B.4 Summary of Q-value Functions for Decentralized Settings

This thesis defined value functions for different settings w.r.t. the assumptions on communication. Also, two types of value functions were identified: expected and
For the general discussion we will consider the immediate reward value functions and make some remarks to provide a coherent perspective. In general, one can interpret the expected reward formulations to correspond to the backwards view (i.e., DP methods): given the joint policy \( q_{t-1}^t \), there is common knowledge of \( b^t \) that lies in the past. This past joint belief can be computed by the agents during execution and acts as a Markovian signal. Given the joint policy \( q_{t-1}^t \) used during the intermediate stages, at stage \( t \) we should select the \( \beta_t^t \) that maximizes the future reward, thereby defining \( V^*_k \) from \( Q^*_k \). To avoid any confusion, Table B.1 shows both \( Q \)-forms used in the thesis.

The immediate reward (IR) formulations are listed in Table B.2. In contrast to ER formulations, these type of value functions express the expected value solely in terms of current joint belief \( b^t \) (induced by some joint action observation history) and future policies \( q_{t+k}^t \). In general, one can interpret the expected reward formulations to correspond to the forward view of Dec-POMDPs (i.e, forward-sweep policy computation etc.): At stage \( t \), there is common knowledge of \( b^{t-k}, q_{t-1}^{t-k} \) which defines a BG for that stage, the solution of which is the maximizing \( \beta_{t+k}^t \). In contrast, the immediate reward formulations correspond to the backwards view (i.e., DP methods): given that \( q_{t+k}^t \) will be used for the first \( k \) stages, \( V^*_k(b^t, q_{t+k}^t) \) specifies the value of joint belief \( b^t \). For instance, \( V^*_h(b^t, q_{t+k}^t) \) gives the expected value of executing \( q_{t+k}^t \) starting from (the state distribution \( b^t \) induced by) \( \tilde{\theta}^t \). That is, \( q_{t+k}^t \) is a joint

<table>
<thead>
<tr>
<th></th>
<th>( V )-form</th>
<th>( Q )-form A</th>
<th>( Q )-form B</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )-SD</td>
<td>( V_{k}^{t,*}(b^{t-k}, q_{t+k}^t) )</td>
<td>( Q_{k}^{t,*}(b^{t-k}, q_{t+k}^t, \beta_{t+k}^t) )</td>
<td>( Q_{k}^{t,*}(b^{t-k}, q_{t+k}^t, \tilde{\theta}<em>{t+k}^t, \beta</em>{t+k}^t) )</td>
</tr>
<tr>
<td>( h )-SD</td>
<td>( V_{h}^{t,*}(b^t, q_{t-h}^t) )</td>
<td>( Q_{h}^{t,*}(b^t, q_{t-h}^t, \beta_{t-h}^t) )</td>
<td>( Q_{h}^{t,*}(b^t, q_{t-h}^t, \tilde{\theta}<em>{t-h}^t, \delta</em>{t-h}^t) )</td>
</tr>
<tr>
<td>1-SD</td>
<td>( V_{1}^{t,*}(b^{t-1}, a^{t-1}) )</td>
<td>( Q_{1}^{t,*}(b^{t-1}, a^{t-1}, \beta_{t-1}^t) )</td>
<td>( Q_{1}^{t,*}(b^{t-1}, a^{t-1}, o^t, \beta_{t-1}^t) )</td>
</tr>
<tr>
<td>0-SD</td>
<td>( V_{0}^{t,*}(b^t, \theta) )</td>
<td>( Q_{0}^{t,*}(b^t, \theta) )</td>
<td>( Q_{0}^{t,*}(b^t, \theta) )</td>
</tr>
</tbody>
</table>

**Table B.1:** Expected reward formulations

<table>
<thead>
<tr>
<th></th>
<th>( V )-form</th>
<th>( Q )-form</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )-SD</td>
<td>( V_{k}^{t,*}(b^t, q_{t+k}^t) )</td>
<td>( Q_{k}^{t,*}(b^t, q_{t+k}^t, \beta_{t+k}^t) )</td>
</tr>
<tr>
<td>( h )-SD</td>
<td>( V_{h}^{t,*}(b^t, q_{t-h}^t) )</td>
<td>( Q_{h}^{t,*}(b^t, q_{t-h}^t, \beta_{t-h}^t, \theta_{t-h}^t) )</td>
</tr>
<tr>
<td>1-SD</td>
<td>( V_{1}^{t,*}(b^t, a^{t}) )</td>
<td>( Q_{1}^{t,*}(b^t, a^{t}, \beta_{t}^t) )</td>
</tr>
<tr>
<td>0-SD</td>
<td>( V_{0}^{t,*}(b^t, \theta) )</td>
<td>( Q_{0}^{t,*}(b^t, \theta) )</td>
</tr>
</tbody>
</table>

**Table B.2:** Immediate reward formulations
sub-tree policy that executes for the remainder of time; stages \( t, \ldots, h - 1 \). This is exactly the relation between the optimal value function and DP for Dec-POMDPs as discussed in Section 3.1.5.1.

It is important to note that \( Q \)-form B, \( Q_{k*}^{t*k}(b^{t-k}, q^{t-k}_{|t|}, \vec{\theta}^{t-k}_{|t|}, \beta^{t}_{|t|}) \), is a function of the joint belief \( b^{t-k} \) (induced by \( \vec{\theta}^{t-k} \)), but not of \( b^{t} \) (as would be induced by \( \vec{\theta}^{t} = (\vec{\theta}^{t-k}, \vec{\theta}^{t}_{|t|}) \)): keeping other arguments the same two histories \( \vec{\theta}^{t}_{|t|}, \vec{\theta}^{t'}_{|t|} \) that would induce the same joint belief \( b^{t} \), can have different values, because \( \vec{\theta}^{t}_{|t|} \) specifies how individual information is distributed. For instance, let us the Dec-Tiger problem with 2-steps delayed communication, \( b^{t-2} \) is uniform and the intermediate joint policy \( q^{t-2}_{|t|} \) specifies only to listen. Now we consider two histories \( \vec{\theta}^{t}_{|t|} = ((\omega_{HL}, \omega_{HL}), (\omega_{HR}, \omega_{HR})) \) while \( \vec{\theta}^{t'}_{|t|} = ((\omega_{HL}, \omega_{HL}), (\omega_{HL}, \omega_{HR})) \). Even though both 2-step joint histories will lead to the same joint belief (a uniform belief over states), the value of the latter will be higher because the agents have a identical individual beliefs. In case of \( \vec{\theta}^{t}_{|t|} = ((\omega_{HL}, \omega_{HL}), (\omega_{HR}, \omega_{HR})) \), however, the agents have a completely different perspective of the world (and are fairly sure that their view is right), therefore the optimal joint policy may specify to open the door in these cases, leading to a lower value for this particular joint history. A more formal argument is given by Oliehoek et al. (2007c).

We end this section by making some specific remarks about the value function for the special cases considered.

The 0-SD setting. Table B.1 and B.2 show that many of the arguments of the general formulation become degenerate under instantaneous communication. E.g., for \( k = 0 \) \( q^{t-k}_{|t|} = () \) is an empty joint sub-tree policy. The result is that the IR and ER formulation are identical: \( V^{t*}_{0} (b^{t}) \) and \( Q^{t*}_{0} (b^{t}, a^{t}) \).

The 1-SD setting. Under 1-step delayed communication, the joint policy followed since \( b^{t-1} \) reduces to a joint action \( q^{t-1}_{|t|} = a^{t-1} \). Also, as discussed in Section 3.3.1, \( Q^{t*}_{1} (b^{t-1}, a^{t-1}, o^{t}, \beta^{t}_{|t|}) \) can be immediately reduced to \( V^{t*}_{1} (b^{t}, a^{t}) \), by noticing that \( b^{t-1}, a^{t-1}, o^{t} \) only influence the value through \( b^{t} \) and that for \( o^{t} \) \( \beta^{t}_{|t|} \) reduces to \( a^{t} = \beta^{t}_{|t|} (o^{t}) \).

A second observation is that \( V^{t*}_{1} (b^{t}, a^{t}) \) has the same arguments as \( Q^{t*}_{0} (b^{t}, a^{t}) \). Because of the apparent similarity, \( V^{t*}_{1} (b^{t}, a^{t}) \) has been denoted as ‘\( Q \)’ in previous work. We still refer to this function as \( Q_{BG} \) when used as a heuristic to stay in line with \( Q_{MDP}, Q_{POMDP} \) naming.

The \( h \)-SD setting. Because ER formulations have their common joint belief at stage 0, \( q^{t-k}_{|t|} \) reduces to a full past joint policy \( \varphi^{t} \). Also, because the joint BG-policies \( \beta^{t}_{|t|} \) are mappings from the entire history, they reduce to joint decision rules \( \delta^{t} \).

In the IR formulation \( Q^{t}_{h} \) is a degenerate form that reduces to \( V^{t}_{h} (b^{t}, q^{t}_{h-t}) \). Also, it is not starred ‘*’ since there is no “continuing optimally afterwards”.