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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
This appendix contains several various proofs that would have interrupted the flow of the text too much. For convenience, the propositions, theorems, lemmas, etc. themselves have been repeated.

D.1 Proofs of Chapter 2

Proof of Theorem 2.1. For a BG with identical payoffs, i.e., \( \forall_i \forall_j \forall \theta \forall a \ u_i(\theta,a) = u_j(\theta,a) \), the solution is given by:

\[
\beta^* = \arg \max_\beta \sum_{\theta \in \Theta} \Pr(\theta)u(\theta,\beta(\theta)), \tag{D.1.1}
\]

where \( \beta(\theta) = (\beta_1(\theta_1),...,\beta_n(\theta_n)) \) is the joint action specified by \( \beta \) for joint type \( \theta \). This solution constitutes a Pareto optimal Nash equilibrium.

Proof. The proof consists of two parts: the first shows that \( \beta^* \) is a Nash equilibrium, the second shows it is Pareto optimal.

Nash Equilibrium Proof. It is clear that \( \beta^* \) satisfying (D.1.1) is a Nash equilibrium by rewriting from the perspective of an arbitrary agent \( i \) as follows:

\[
\beta^*_i = \arg \max_{\beta_i} \left[ \max_{\beta_{\neq i}} \sum_{\theta \in \Theta} \Pr(\theta)u(\theta,\beta(\theta)) \right],
\]

\[
= \arg \max_{\beta_i} \left[ \max_{\beta_{\neq i}} \sum_{\theta_i} \sum_{\theta_{\neq i}} \Pr(\theta_i | \theta_i) \left\{ \sum_{\theta_{\neq i}} \Pr(\langle \theta_i, \theta_{\neq i} \rangle) u(\theta,\beta(\theta)) \right\} \right],
\]

\[
= \arg \max_{\beta_i} \left[ \max_{\beta_{\neq i}} \sum_{\theta_i} \Pr(\theta_i) \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i} | \theta_i) u(\theta,\beta(\theta)) \right],
\]

\[
= \arg \max_{\beta_i} \sum_{\theta_i} \Pr(\theta_i) \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i} | \theta_i) u(\langle \theta_i, \theta_{\neq i} \rangle, \langle \beta_i(\theta_i), \beta^*_{\neq i}(\theta_{\neq i}) \rangle),
\]

185
which means that $\beta^*_i$ is a best response for $\beta^*_{\neq i}$. Since no special assumptions were made on $i$, it follows that $\beta^*$ is a Nash equilibrium.

**Pareto Optimality Proof.** Let us write $V_{\theta_i}(a_i, \beta_{\neq i})$ for the payoff agent $i$ expects for $\theta_i$ when performing $a_i$ when the other agents use policy profile $\beta_{\neq i}$. We have that

$$V_{\theta_i}(a_i, \beta_{\neq i}) = \sum_{\theta_{\neq i}} \Pr(\theta_{\neq i}; \theta_i)u(\langle \theta_i, \theta_{\neq i} \rangle, \langle a_i, \beta_{\neq i}(\theta_{\neq i}) \rangle).$$

Now, a joint policy $\beta^*$ satisfying (D.1.1) is not Pareto optimal if and only if there is another Nash equilibrium $\beta'$ that attains at least the same payoff for all agents $i$ and for all types $\theta_i$ and strictly more for at least one agent and type. Formally $\beta^*$ is not Pareto optimal when $\exists \beta'$ such that, $\forall_i \forall_{\theta_i}$

$$V_{\theta_i}(\beta_i^*(\theta_i), \beta^*_{\neq i}) \leq V_{\theta_i}(\beta_i'(\theta_i), \beta^*_{\neq i}) \land \exists \theta_i \exists \theta_{\neq i} V_{\theta_i}(\beta_i^*(\theta_i), \beta^*_{\neq i}) < V_{\theta_i}(\beta_i'(\theta_i), \beta^*_{\neq i}).$$

(D.1.2)

We prove that no such $\beta'$ can exist by contradiction. Suppose that $\beta' = \langle \beta_i', \beta^*_{\neq i} \rangle$ is a NE such that (D.1.2) holds (and thus $\beta^*$ is not Pareto optimal). Because $\beta^*$ satisfies (D.1.1) we know that:

$$\sum_{\theta \in \Theta} \Pr(\theta)u(\theta, \beta^*(\theta)) \geq \sum_{\theta \in \Theta} \Pr(\theta)u(\theta, \beta'(\theta)),$$

and therefore, for all agents $i$

$$\Pr(\theta_{i, 1})V_{\theta_{i, 1}}(\beta_i^*(\theta_i), \beta^*_{\neq i}) + \ldots + \Pr(\theta_{i, |\Theta_i|})V_{\theta_{i, |\Theta_i|}}(\beta_i^*(\theta_i), \beta^*_{\neq i}) \geq \Pr(\theta_{i, 1})V_{\theta_{i, 1}}(\beta_i'(\theta_i), \beta^*_{\neq i}) + \ldots + \Pr(\theta_{i, |\Theta_i|})V_{\theta_{i, |\Theta_i|}}(\beta_i'(\theta_i), \beta^*_{\neq i})$$

holds. However, by assumption that $\beta'$ satisfies (D.1.2) we get that

$$\exists_j \quad V_{\theta_{i, j}}(\beta_i^*(\theta_i), \beta^*_{\neq i}) < V_{\theta_{i, j}}(\beta_i'(\theta_i), \beta^*_{\neq i}).$$

Therefore it must be that

$$\sum_{k \neq j} \Pr(\theta_{i, k})V_{\theta_{i, k}}(\beta_i^*(\theta_i), \beta^*_{\neq i}) > \sum_{k \neq j} \Pr(\theta_{i, k})V_{\theta_{i, k}}(\beta_i'(\theta_i), \beta^*_{\neq i}),$$

and thus that

$$\exists_k \quad V_{\theta_{i, k}}(\beta_i^*(\theta_i), \beta^*_{\neq i}) > V_{\theta_{i, k}}(\beta_i'(\theta_i), \beta^*_{\neq i}),$$

contradicting the assumption that $\beta'$ satisfies (D.1.2). \hfill \Box

### D.2 Proofs of Chapter 3

**Proof of 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:

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

(D.2.1)
where \( \tilde{\theta}^t = (\tilde{\theta}^t, \tilde{a}^t) \), where \( \pi^*(\tilde{\theta}^t) = \pi^*\tilde{\theta}^t \) denotes the joint action \( \pi^* \) specifies for \( \tilde{\theta}^t \), and where

\[
Q_{\pi^*}(\tilde{\theta}^t, a) = R(\tilde{\theta}^t, a) + \sum_{\tilde{\theta}^{t+1} \in \tilde{\Theta}} \Pr(\tilde{\theta}^{t+1} | \tilde{\theta}^t, a) Q_{\pi^*}(\tilde{\theta}^{t+1}, \pi^*(\tilde{\theta}^{t+1})) \tag{D.2.2}
\]

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

**Proof.** By filling out (2.5.5) for an optimal pure joint policy \( \pi^* \), we obtain its expected cumulative reward as the summation of \( E[R(s^t, a^t) | \pi^*] \) the expected rewards it yields for each time step:

\[
V(\pi^*) = \sum_{t=0}^{h-1} E[R(s^t, a^t) | \pi^*] = \sum_{t=0}^{h-1} \sum_{\tilde{\theta}^t \in \tilde{\Theta}} \Pr(\tilde{\theta}^t | \pi^*, b^0) R(\tilde{\theta}^t, \pi^*(\tilde{\theta}^t)). \tag{D.2.3}
\]

In this equation, \( \Pr(\tilde{\theta}^t | \pi^*, b^0) \) is given by (3.1.3). As a result, the influence of \( \pi^* \) on \( \Pr(\tilde{\theta}^t | \pi^*, b^0) \) is only through \( C \). I.e., \( \pi^* \) is only used to ‘filter out’ inconsistent histories. Therefore we can write:

\[
E[R(s^t, a^t) | \pi^*] = \sum_{\tilde{\theta}^t \in \tilde{\Theta}_{\pi^*}} \Pr(\tilde{\theta}^t | b^0) R(\tilde{\theta}^t, \pi^*(\tilde{\theta}^t)), \tag{D.2.4}
\]

where \( \Pr(\tilde{\theta}^t | b^0) \) is given by directly taking the marginal of (3.1.4). Now, let us define the value starting from time step \( t \):

\[
V^t(\pi^*) = E[R(s^t, a^t) | \pi^*] + V^{t+1}(\pi^*)
= \sum_{\tilde{\theta}^t \in \tilde{\Theta}_{\pi^*}} \Pr(\tilde{\theta}^t | b^0) R(\tilde{\theta}^t, \pi^*(\tilde{\theta}^t)) + V^{t+1}(\pi^*). \tag{D.2.5}
\]

For the last time step \( h - 1 \) there is no expected future reward, so we get:

\[
V^{h-1}(\pi^*) = \sum_{\tilde{\theta}^{h-1} \in \tilde{\Theta}^{h-1}_{\pi^*}} \Pr(\tilde{\theta}^{h-1} | b^0) R(\tilde{\theta}^{h-1}, \pi^*(\tilde{\theta}^{h-1})) Q_{\pi^*}(\tilde{\theta}^{h-1}, \pi^*(\tilde{\theta}^{h-1})). \tag{D.2.6}
\]

For time step \( h - 2 \) this becomes:

\[
V^{h-2}(\pi^*) \equiv E[R(s^{h-2}, a^{h-2}) | \pi^*] + V^{h-1}(\pi^*) = \sum_{\tilde{\theta}^{h-2} \in \tilde{\Theta}^{h-2}_{\pi^*}} \Pr(\tilde{\theta}^{h-2} | b^0) R(\tilde{\theta}^{h-2}, \pi^*(\tilde{\theta}^{h-2})) + \sum_{\tilde{\theta}^{h-1} \in \tilde{\Theta}^{h-1}_{\pi^*}} \Pr(\tilde{\theta}^{h-1} | b^0) Q_{\pi^*}(\tilde{\theta}^{h-1}, \pi^*(\tilde{\theta}^{h-1})). \tag{D.2.7}
\]

Because \( \Pr(\tilde{\theta}^{h-1}) = \Pr(\tilde{\theta}^{h-2}) \Pr(\tilde{\theta}^{h-1} | \tilde{\theta}^{h-2}, \pi^*(\tilde{\theta}^{h-2})) \), (D.2.7) can be rewritten to:

\[
V^{h-2}(\pi^*) = \sum_{\tilde{\theta}^{h-2} \in \tilde{\Theta}^{h-2}_{\pi^*}} \Pr(\tilde{\theta}^{h-2} | b^0) Q_{\pi^*}(\tilde{\theta}^{h-2}, \pi^*(\tilde{\theta}^{h-2})). \tag{D.2.8}
\]
Proof of 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:

$$V^t_{k,*}(b^{t-k}, q^{t-k}) = \max_{\beta^t} Q^t_{k,*}(b^{t-k}, q^{t-k}, \beta^t).$$  \hspace{1cm} (D.2.10)

$$Q^t_{k,*}(b^{t-k}, q^{t-k}, \beta^t) = \sum_{\bar{\theta}^t_{[k]}} \Pr(\bar{\theta}^t_{[k]}|b^{t-k}, q^{t-k}) Q^t_{k,*}(b^{t-k}, q^{t-k}, \bar{\theta}^t_{[k]}, \beta^t).$$  \hspace{1cm} (D.2.11)

$$Q^t_{k,*}(b^{t-k}, q^{t-k}, \bar{\theta}^t_{[k]}, \beta^t) = R(b^t, \beta^t(\bar{\theta}^t_{[k]})) + \sum_{\bar{o}^{t+1}} \Pr(o^{t+1}|b^t, \beta^t(\bar{\theta}^t_{[k]})) Q^{t+1,*}_{k,\pi}(b^{t-k+1}, q^{t-k+1}, \bar{\theta}^{t+1}_{[k]}, \beta^{t+1,*}).$$  \hspace{1cm} (D.2.12)

where $b^t$ results from $b^{t-k}, \bar{\theta}^t_{[k]}$, and

$$\beta^{t+1,*} = \arg\max_{\beta^{t+1}} \sum_{\bar{\theta}^{t+1}} \Pr(\bar{\theta}^{t+1}_{[k]}|b^{t-k+1}, q^{t-k+1}) Q^{t+1,*}_{k,\pi}(b^{t-k+1}, q^{t-k+1}, \bar{\theta}^{t+1}_{[k]}, \beta^{t+1}).$$  \hspace{1cm} (D.2.13)

Proof. We will show that

$$Q^t_{k,*}(b^{t-k}, q^{t-k}, \beta^t)$$

$$= E\left[R(s^t, a^t)|b^{t-k}, q^{t-k}, \beta^t\right] + E\left[V^{t+1,*}_{k,\pi}(b^{t-k+1}, q^{t-k+1}) \mid b^{t-k}, q^{t-k}, \beta^t\right].$$  \hspace{1cm} (D.2.14)

and that the other equations are consistent with this definition. This means that, for the last stage (D.2.10) maximizes the expected reward and therefore is optimal, optimality for other stages follows immediately by induction.

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

$$+ \left[ \sum_{\bar{o}^{t-k+1}} \Pr(o^{t-k+1}|b^{t-k}, q^{t-k}) V^{t+1,*}_{k,\pi}(b^{t-k+1}, q^{t-k+1}) \right].$$  \hspace{1cm} (D.2.15)
where \( q^{t-k+1} = (q^{t-k} \circ \beta^t)\| q^{t-k+1} \), and \( b^{t-k+1} \) results from \( b^{t-k} \) via the joint action \( a^{t-k} \) (specified by \( q^{t-k} \)) and \( o^{t-k+1} \).

By introducing \( \beta^{t+1} \) as in (D.2.13), \( V_k^{t+1,*} \) can be replaced with

\[
V_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}) = \max_{\beta^{t+1}} \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k+1}, q^{t-k+1}) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1}) \quad (D.2.16)
\]

\[
= \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k+1}, q^{t-k+1}) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \quad (D.2.17)
\]

The result of this replacement is

\[
Q_k^{t,*}(b^{t-k}, q^{t-k}, \beta^t) = \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k}, q^{t-k}) R(b^t, \beta^t(\theta_{|k|})) + \left[ \sum_{o^{t-k+1}} \Pr(o^{t-k+1} | b^{t-k}, q^{t-k}, \beta^t) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \right] \]

\[
Q_k^{t,*}(b^{t-k}, q^{t-k}, \beta^t) = \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k}, q^{t-k}) R(b^t, \beta^t(\theta_{|k|})) + \left[ \sum_{\theta_{|k+1|}^{t+1}} \Pr(\theta_{|k+1|}^{t+1} | b^{t-k}, q^{t-k}, \beta^t) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \right] \]

\[
Q_k^{t,*}(b^{t-k}, q^{t-k}, \beta^t) = \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k}, q^{t-k}) R(b^t, \beta^t(\theta_{|k|})) + \left[ \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k}, q^{t-k}) \sum_{o^{t+1}} \Pr(o^{t+1} | b^{t}, \beta^t(\theta_{|k|})) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \right] \]

\[
Q_k^{t,*}(b^{t-k}, q^{t-k}, \beta^t) = \sum_{\theta_{|k|}} \Pr(\theta_{|k|} | b^{t-k}, q^{t-k}) R(b^t, \beta^t(\theta_{|k|})) + \left[ \sum_{o^{t+1}} \Pr(o^{t+1} | b^{t}, \beta^t(\theta_{|k|})) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \right] \]

and thus, finally,

\[
Q_k^{t,*}(b^{t-k}, q^{t-k}, \theta_{|k|}, \beta^t) = R(b^t, \beta^t(\theta_{|k|})) + \sum_{o^{t+1}} \Pr(o^{t+1} | b^t, \beta^t(\theta_{|k|})) Q_k^{t+1,*}(b^{t-k+1}, q^{t-k+1}, \theta_{|k|}, \beta^{t+1,*}) \quad (D.2.18)
\]

concluding the proof.
Proof of 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 $q_{k+1}^{t-k-1}$ followed during stages $t-k, \ldots, t-1$ is no less under $k$-steps communication delay, than under $(k+1)$-steps delay. That is

$$\forall t \forall b^{t-k-1} \forall q_{k+1}^{t-k-1} \quad E[V_{k}^{t,*}(b^{t-k}, q_{k}^{t-k}) | b^{t-k-1}, q_{k+1}^{t-k-1}] \geq V_{k+1}^{t,*}(b^{t-k}, q_{k+1}^{t-k-1}).$$

(D.2.19)

Proof. The proof is by induction. The base case is that (D.2.19) holds for stage $t = h - 1$. The induction hypothesis is that (D.2.19) holds for some stage $t + 1$

$$E[V_{k+1}^{t+1,*}(b^{t-k+1}, q_{k+1}^{t-k+1}) | b^{t-k}, q_{k+1}^{t-k}] \geq V_{k+1}^{t+1,*}(b^{t-k}, q_{k+1}^{t-k}).$$

(D.2.20)

We now need to show that (D.2.19) holds given (D.2.20). Assuming an arbitrary $t < h - 1$, $\bar{\theta}^{t-k-1}$ and $q_{k+1}^{t-k-1}$, the left side of (D.2.19) can be rewritten as follows

$$E[V_{k}^{t,*}(b^{t-k}, q_{k}^{t-k}) | b^{t-k-1}, q_{k+1}^{t-k-1}]$$

(D.2.21)

$$= E \left[ \max_{\beta_{k}} Q_{k}^{t,*}(b^{t-k}, q_{k}^{t-k}) | b^{t-k-1}, q_{k+1}^{t-k-1} \right]$$

(D.2.22)

$$= E \left[ \max_{\beta_{k}} \left( E[R(s, a) | b^{t-k}, q_{k}^{t-k}, \beta_{k}] + E[V_{k+1}^{t+1,*}(b^{t-k+1}, q_{k+1}^{t-k+1}) | b^{t-k}, q_{k+1}^{t-k}, \beta_{k}] \right) \right]$$

(D.2.23)

Note that $q_{k+1}^{t-k}, \beta_{k}$ together form $q_{k+1}^{t-k}$. I.e., $q_{k+1}^{t-k} = (q_{k}^{t-k} \circ \beta_{k})$. Now the induction hypothesis can be applied:

$$= E \left[ \max_{\beta_{k}} \left( E[R(s, a) | b^{t-k}, q_{k}^{t-k}, \beta_{k}] + E[V_{k+1}^{t+1,*}(b^{t-k+1}, q_{k+1}^{t-k+1}) | b^{t-k}, q_{k+1}^{t-k}, \beta_{k}] \right) \right]$$

(D.2.24)

Now we make the outer expectation over $a^{t-k}$ explicit. In particular $b^{t-k}$ depends on $a^{t-k}$ and $q_{k+1}^{t-k} = q_{k+1}^{t-k} \downarrow_{a^{t-k}}$, leading to

$$= \sum_{a^{t-k}} \Pr(o^{t-k} | b^{t-k-1}, q_{k+1}^{t-k-1}) \max_{\beta_{k}} \left[ E[R(s, a) | b^{t-k}, q_{k+1}^{t-k-1} \downarrow_{a^{t-k}}, \beta_{k}] + V_{k+1}^{t+1,*}(b^{t-k}, q_{k+1}^{t-k-1} \downarrow_{a^{t-k}} \circ \beta_{k}) \right]$$

(D.2.24)
where $\beta'_{[k]}$ is that length-$k$ joint BG policies that is selected (is maximizing) for $\mathbf{o}^{t-k}$. Let us combine these selected policies in one ‘First-Joint-Observation policy’ term: $\beta_{FJO,t}^{i} = \langle \beta_{FJO,t}^{i}, |i| \in D \rangle$ where an individual policy maps $\mathbf{o}^{t-k}$ to $\beta_{[i]}^{t}$ the individual length-$k$ BG policy $\beta_{FJO,t}^{i} (\mathbf{o}^{t-k}) = \beta_{[i]}^{t}$ and thus

\[
\beta_{FJO,t}^{i} : \mathbf{o}^{t-k} \times \bar{O}_{t}^{k} \rightarrow A_{i}.
\]

We also write $\beta_{FJO,t}^{i} \|_{\mathbf{o}^{t-k}} = \beta_{[i]}^{t}$. Using this notation, we can rewrite to

\[
= \max_{\beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}}} \sum \Pr(\mathbf{a}^{t-k}|b^{t-k-1}, q_{k+1}^{t-k-1}) \left[ E[R(s^{t}, a^{t}) | b^{t-k}, q_{k+1}^{t-k}, D.2.25 \rangle] + V^{t+1}_{k+1} (b^{t-k}, q_{k+1}^{t-k}) \right]
\]

Because the set of joint policies $\beta_{FJO,t}^{i}$ is a strict superset of the set of all possible $\beta_{[k+1]}^{t}$, we get

\[
\geq \max_{\beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}}} \sum \Pr(\mathbf{a}^{t-k}|b^{t-k-1}, q_{k+1}^{t-k-1}) \left[ E[R(s^{t}, a^{t}) | b^{t-k}, q_{k+1}^{t-k}, \beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}}] + V^{t+1}_{k+1} (b^{t-k}, q_{k+1}^{t-k}) \right]
\]

Now, using $\langle q_{k+1}^{t-k-1} \|_{\mathbf{o}^{t-k}} \beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}} \rangle = \langle q_{k+1}^{t-k-1} \|_{\beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}}} \rangle$, and

\[
E[R(s^{t}, a^{t}) | b^{t-k}, q_{k+1}^{t-k}, \beta_{[k+1]}^{t}] = \sum \Pr(\theta_{[k]}^{t} | b^{t-k}, q_{k+1}^{t-k}) R(b^{t}, \beta_{[k]}^{t} (\theta_{[k]}^{t}))
\]

we write

\[
= \max_{\beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}}} \sum \Pr(\mathbf{a}^{t-k}|b^{t-k-1}, q_{k+1}^{t-k-1}) \left[ \sum \Pr(\theta_{[k]}^{t} | b^{t-k}, q_{k+1}^{t-k} \|_{\mathbf{o}^{t-k}}) \right]
\]

\[
R(b^{t}, \beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}} (\theta_{[k]}^{t})) + V^{t+1}_{k+1} (b^{t-k}, q_{k+1}^{t-k}) \left[ E[R(s^{t}, a^{t}) | b^{t-k}, q_{k+1}^{t-k}, \beta_{[k+1]}^{t} \|_{\mathbf{o}^{t-k}} \rangle] \right]
\]
writing $q_{|k+1|}^{t-k} = (q_{|k+1|}^{t-k-1} \circ \beta_{|k+1|}^t) \|_{o^{t-k}}$ we further reduce

$$\max_{\beta_{|k+1|}} \left[ E[R(s^t, a^t) \mid b^{t-k-1}, q_{|k+1|}^{t-k-1} \circ \beta_{|k+1|}^t] ight]$$

$$+ E[V_{k+1}^{t+1,*}(b^{t-k}, q_{|k+1|}^{t-k}) \mid b^{t-k}, q_{|k+1|}^{t-k}]$$

$$= \max_{\beta_{|k+1|}} Q_{k+1}^{t,*}(b^{t-k}, q_{|k+1|}^{t-k} \circ \beta_{|k+1|})$$

$$= V_{k+1}^{t,*}(b^{t-k}, q_{|k+1|}^{t-k})$$

which proves the induction step. The prove is completed by the base case, which is given by the last stage. I.e., we need to show that for an arbitrarily chosen $b^{h-k-2}$ (corresponding to some $\tilde{\theta}^{h-k-2}$) and $q_{|k+1|}^{h-k-2}$

$$E[V_{k}^{h-1,*}(b^{h-1-k}, q_{|k|}^{h-1-k}) \mid b^{h-k}, q_{|k+1|}^{h-k-2}] \geq V_{k+1}^{h-1,*}(b^{h-1-k}, q_{|k+1|}^{h-k-2}).$$

Starting from the left side, we make the expectation explicit

$$E[V_{k}^{h-1,*}(b^{h-1-k}, q_{|k|}^{h-1-k}) \mid b^{h-k}, q_{|k+1|}^{h-k-2}] =$$

$$\sum_{o^{h-1-k}} \Pr(o^{h-1-k} \mid b^{h-k}, q_{|k+1|}^{h-k-2}) V_{k}^{h-1,*}(b^{h-1-k}, q_{|k+1|}^{h-k-2} \|_{o^{h-1-k}})$$

(D.2.33)

because we consider the last stage $V_{k}^{h-1,*}$ only consists of the expected immediate reward

$$V_{k}^{h-1,*}(b^{h-1-k}, q_{|k+1|}^{h-k-2} \|_{o^{h-1-k}})$$

$$= \max_{\beta_{|k|}} E[R(s^{h-1}, a^{h-1}) \mid b^{h-k-1}, q_{|k+1|}^{h-k-2} \|_{o^{h-1-k}}, \beta_{|k|}^{h-1}]$$

$$= \max_{\beta_{|k|}} \sum_{\tilde{\theta}_{|k|}^{h-1}} \Pr(\tilde{\theta}_{|k|}^{h-1} \mid b^{h-k-1}, q_{|k+1|}^{h-k-2} \|_{o^{h-1-k}}) R(b^{h-1}, \beta_{|k|}^{h-1}(\tilde{\theta}_{|k|}^{h-1}))$$

(D.2.36)

Thus (D.2.33) equals

$$\sum_{o^{h-1-k}} \Pr(o^{h-1-k} \mid b^{h-k}, q_{|k+1|}^{h-k-2}) \max_{\beta_{|k|}^{h-1}} \sum_{\tilde{\theta}_{|k|}^{h-1}} \Pr(\tilde{\theta}_{|k|}^{h-1} \mid b^{h-k-1}, q_{|k+1|}^{h-k-2} \|_{o^{h-1-k}}) R(b^{h-1}, \beta_{|k|}^{h-1}(\tilde{\theta}_{|k|}^{h-1}))$$

(D.2.37)
Again using FJO policies, we can write
\[
\begin{align*}
&= \max_{\beta^{FJO,h-1}_{[k+1]}} \sum_{o^{h-1-k}} \Pr(o^{h-1-k}|b^{h-k-2},q_{[k+1]}) \\
&= \sum_{\theta^{h-1}_{|k|}} \Pr(\theta^{h-1}_{|k|}|b^{h-k-1},q_{[k+1]}) R(b^{h-1},\beta_{[k+1]}) (\theta^{h-1}_{|k|})) \\
&= \sum_{\theta^{h-1}_{|k|}} \Pr(\theta^{h-1}_{|k|}|b^{h-k-1},q_{[k+1]}) R(b^{h-1},\beta^{h-1}_{[k+1]}(\theta^{h-1}_{|k|})) \\
&= V^{h-1,*}(b^{h-k-2},q_{[k+1]}), \quad (D.2.39)
\end{align*}
\]

because the set of FJO policies $\beta^{FJO,h-1}_{[k+1]}$ is a strict superset of the set of $\beta^{h-1}_{[k+1]}$. Taking together the summations in (D.2.39) yields
\[
\begin{align*}
&= \max_{\beta^{h-1}_{[k+1]}} \sum_{\theta^{h-1}_{|k|}} \Pr(\theta^{h-1}_{|k|}|b^{h-k-1},q_{[k+1]}) R(b^{h-1},\beta^{h-1}_{[k+1]}(\theta^{h-1}_{|k|})) \\
&= V^{h-1,*}(b^{h-k-2},q_{[k+1]}), \quad (D.2.40)
\end{align*}
\]

which proves the base case.

\section*{D.3 Proofs of Chapter 5}

\textbf{Proof of Theorem 5.1} (Decomposition of $V^t(\pi)$). Given an additively factored immediate reward function, the value $V^t(\pi)$ of a finite-horizon factored Dec-POMDP is decomposable for any $t$. That is, for any joint policy $\pi$ the value function is factored. $V^t(\pi)$ is defined as
\[
V^t(\pi) = \sum_{e \in E} V^{e,t}(\pi) = \sum_{e \in E} \sum_{x_t^{\theta_{A_e}} \in X_e} \sum_{\theta_{A_e}} \Pr(x_t^{\theta_{A_e}}, \theta_{A_e} | b^0, \pi) Q^e(\pi)(x_t^{\theta_{A_e}}, \theta_{A_e}, \pi_{A_e}(\theta_{A_e}))
\]
\[
(D.3.1)
\]
where, using shorthand notation $\Gamma^X_e = \Gamma^X(x_t^{f+1} \cup o_t^{f+1})$ and $\Gamma^A_e = \Gamma^A(x_t^{f+1} \cup o_t^{f+1})$ to denote the backup scopes and $X_e = X(R_e) \cup \Gamma^X_e$ and $A_e = A(R_e) \cup \Gamma^A_e$ to denote the scopes of $Q^e_{\pi,t}$,
\[
Q^e(\pi)(x_t^{\theta_{A_e}}, \theta_{A_e}, a_{A_e}) = R^e(x_t^{A_e}, a_e) + \sum_{x_{t+1}^{f+1}, o_{t+1}^{f+1}} \Pr(x_{t+1}^{f+1}, o_{t+1}^{f+1} | x_t^{f+1}, a_t^{f+1}) Q^e(\pi)(x_{t+1}^{f+1}, \theta_{A_e}, \pi_{A_e}(\theta_{A_e})). \quad (D.3.2)
\]

\textbf{Proof.} This proof assumes $\pi$ is a pure joint policy, but can be generalized to stochastic policies. Per induction hypothesis, we assume the next-stage value function is decomposable:
\[
V^{t+1}(\pi) = \sum_{e \in E} V^{e,t+1}(\pi) \quad (D.3.3)
\]
and that each $V^{e,t+1}$ can be written using Q-value functions:

$$V^{e,t+1}(\pi) = \sum_{x_{t+1}^e} \sum_{\theta_{t+1}^e} \Pr(x_{t+1}^e, \theta_{t+1}^e | b^0, \pi) Q^e_{\pi}(x_{t+1}^e, \theta_{t+1}^e, \pi_{\theta_e}((\theta_{t+1}^e)))$$  \hspace{1cm} (D.3.4)

with $X_{t}^{\prime}, A_{t}$ the scopes of $Q^e_{\pi}$ for stage $t+1$. Now, we can show that $V^t$ is decomposable, although the scope of $V^t$ might grow. Per definition we have that

$$V^t(\pi) = E[R(s^t, a) | b^0, \pi] + V^{t+1}(\pi) = \sum_{s^t, \theta^t} \Pr(s^t, \theta^t | b^0, \pi) R(s^t, \pi(\theta^t)) + \sum_{e \in E} V^{e,t+1}(\pi)$$  \hspace{1cm} (D.3.5)

By applying the definition of the additive immediate reward function and using the induction hypothesis we get

$$V^t(\pi) = \sum_{s^t, \theta^t} \Pr(s^t, \theta^t | b^0, \pi) \sum_{e \in E} R^e(x_{t}^e, \pi_e((\theta_{t}^e))) +$$

$$\sum_{e \in E} \sum_{x_{t+1}^e} \sum_{\theta_{t+1}^e} \Pr(x_{t+1}^e, \theta_{t+1}^e | b^0, \pi) Q^e_{\pi}(x_{t+1}^e, \theta_{t+1}^e, \pi_{\theta_e}((\theta_{t+1}^e)))$$  \hspace{1cm} (D.3.6)

We can decompose $\Pr(x_{t+1}^{e} | b^0, \pi)$ in a fashion similar to (2.5.6) and obtain

$$V^t(\pi) = \sum_{s^t, \theta^t} \Pr(s^t, \theta^t | b^0, \pi) \sum_{e \in E} R^e(x_{t}^e, \pi_e((\theta_{t}^e))) +$$

$$\sum_{s^t, \theta^t} \Pr(s^t, \theta^t | b^0, \pi) \sum_{e \in E} \sum_{o_{t+1}^{e}} \Pr(o_{t+1}^{e} | s^t, \pi(\theta^t)) Q^e_{\pi}(x_{t+1}^{e}, \theta_{t+1}^{e}, \pi_{\theta_e}((\theta_{t+1}^{e})))$$  \hspace{1cm} (D.3.7)

and thus yield

$$V^t(\pi) = \sum_{s^t, \theta^t} \Pr(s^t, \theta^t | b^0, \pi) \sum_{e \in E} R^e(x_{t}^e, \pi_e((\theta_{t}^e))) +$$

$$\sum_{x_{t+1}^{e}, o_{t+1}^{e}} \Pr(x_{t+1}^{e}, o_{t+1}^{e} | s^t, \pi(\theta^t)) Q^e_{\pi}(x_{t+1}^{e}, \theta_{t+1}^{e}, \pi_{\theta_e}((\theta_{t+1}^{e})))$$  \hspace{1cm} (D.3.8)

where the bracketed part acts as a Q-value function for stage $t$. When there is no independence this value depends on the full $s^t, \pi(\theta^t)$ and thus the scope of this value function includes all states factors and agents. However, we can exploit any independence that does hold. In particular, we know that the probability of a pair $x_{t+1}^{e}, o_{t+1}^{e}$ is influenced only by the parents in the DBN: the state factors with indices $\Gamma_{X}^{\alpha}(x_{t+1}^{e} \cup o_{t+1}^{e})$ and the action nodes of agents $\Gamma_{A}^{\alpha}(x_{t+1}^{e} \cup o_{t+1}^{e})$. We use
shorthand notation $\Gamma^X$ and $\Gamma^A$ to derive

$$V^t(\pi) = \sum_{s^t, \bar{\theta}^t} \Pr(s^t, \bar{\theta}^t | b^0, \pi) \sum_{e \in E} \left[ R^e(x^t_e, \pi^e(\bar{\theta}^t_{e})) + \sum_{x^t_{\bar{e}}} \sum_{o^t_{\bar{e}}} \Pr(x^{t+1}_{\bar{e}}, o^{t+1}_{\bar{e}} | x^t_{\bar{e}}, \pi_{\bar{e}} \Gamma(\bar{\theta}^t_{\bar{e}})) Q^e_{\pi}(x^{t+1}_{\bar{e}}, \bar{\theta}^{t+1}_{\bar{e}}, \pi_{\bar{e}} (\bar{\theta}^{t+1}_{\bar{e}})) \right]. \tag{D.3.9}$$

We write $X_e = X(R^e) \cup \Gamma^X$, $\Lambda_e = \Lambda(R^e) \cup \Gamma^A \cup A'_{\Lambda}$ for the union of scopes of the entire bracketed part and swap the summations

$$V^t(\pi) = \sum_{e \in E} \sum_{x^t_{\bar{e}}} \sum_{\bar{\theta}^t_{\bar{e}}} \Pr(x^t_{\bar{e}}, \bar{\theta}^t_{\bar{e}} | b^0, \pi) \left[ R^e(x^t_e, \pi^e(\bar{\theta}^t_{e})) + \sum_{x^{t+1}_{\bar{e}}} \sum_{o^{t+1}_{\bar{e}}} \Pr(x^{t+1}_{\bar{e}}, o^{t+1}_{\bar{e}} | x^t_{\bar{e}}, \pi_{\bar{e}} \Gamma(\bar{\theta}^t_{\bar{e}})) Q^e_{\pi}(x^{t+1}_{\bar{e}}, \bar{\theta}^{t+1}_{\bar{e}}, \pi_{\bar{e}} (\bar{\theta}^{t+1}_{\bar{e}})) \right]. \tag{D.3.10}$$

At this point we have derived (D.3.1) and (D.3.2). The last stage as treated in Lemma 5.1 forms the base case and thus completes the proof. \hfill \Box