do not include full proof of this claim. Intuitively it holds because each agent does not receive any information about the other agents’ local histories due to transition independence. Therefore, we can represent agent 1’s policy on the last step as \( \sigma^*_{T-1}(h^1_{T-1}) = \arg \max_{a^1} \sum_{h^2_{T-1}} P(h^2_{T-1}) \cdot R(s^1, a^1, \sigma^2_{T-1}(h^2_{T-1})) \) which no longer depends on the history \( h^1_{T-1} \). Therefore, the policy on the last step for either agent does not depend on history.

This allows us to define the value function on the last step as \( v_{T-1}(s, \sigma^1_{T-1}(z^1_{T-1}), \sigma^2_{T-1}(z^2_{T-1})) \).

Then for the induction step, we can show that if the policy at step \( \tau + 1 \) does not depend on history, then the policy at step \( \tau \) also does not depend on its local history. Again, we show this from agent 1’s perspective.

Agent 1’s policy on step \( \tau \) can be represented by:

\[
\sigma^1_{\tau}(h^1_{\tau}) = \arg \max_{a^1} \sum_{h^2_{\tau}} P(h^2_{\tau} | h^1_{\tau}) \cdot v_{\tau+1}(s, a^1, \sigma^2_{\tau}(h^2_{\tau})),
\]

where the value function \( v_{\tau+1} \) is assumed to not depend on history. We can again show that \( P(h^2_{\tau} | h^1_{\tau}) = P(h^2_{\tau}) \) because of transition independence and represent agent 1’s policy on step \( \tau \) as:

\[
\sigma^1_{\tau}(h^1_{\tau}) = \arg \max_{a^1} \sum_{h^2_{\tau}} P(h^2_{\tau}) \cdot v_{\tau+1}(s, a^1, \sigma^2_{\tau}(h^2_{\tau})),
\]

which no longer depends on the local history \( h^1_{\tau} \).

Therefore, the policy of either agent does not depend on local history for any step of the problem.

We now establish the sufficient statistic for the selection of decentralized Markov decision rules.

**Theorem 2 (Sufficient Statistic)** The state occupancy is a sufficient statistic for decentralized Markov decision rules.

**Proof** We build upon the proof of the optimality of decentralized Markov policies in Theorem 1. We note that an optimal decentralized Markov policy starting in \( \eta_0 \) is given by:

\[
\pi^* = \arg \max_{\pi} \sum_{s} \sum_{a_{\tau-1}, s_{\tau-1}} P(h_{\tau} | \sigma_{0:\tau-1}, \eta_0) \cdot r(s_{\tau}, \sigma_{\tau}[s_{\tau}])
\]

The substitution of \( h_{\tau} \) by \( (h_{\tau-1}, a_{\tau-1}, s_{\tau}) \) plus the sum over all pairs \( (h_{\tau-1}, a_{\tau-1}) \) yields

\[
\pi^* = \arg \max_{\pi} \sum_{s} \sum_{a_{\tau-1}, s_{\tau-1}} P(s_{\tau} | \sigma_{0:\tau-1}, \eta_0) \cdot r(s_{\tau}, \sigma_{\tau}[s_{\tau}]),
\]

We denote \( \eta^*_{\tau} = P(s_{\tau} | \sigma_{0:\tau-1}, \eta_0) \) the state occupancy distribution that decentralized Markov policy \( \pi \) produced at horizon \( \tau \). And hence,

\[
\pi^* = \arg \max_{\pi} \sum_{s} \sum_{s_{\tau} \in S} \eta^*_{\tau}(s_{\tau}) \cdot r(s_{\tau}, \sigma_{\tau}[s_{\tau}])
\]

So, state occupancy \( \eta^*_{\tau} \) summarizes all possible joint action-observation histories \( h_{\tau} \) decentralized Markov policy \( \pi \) produced at horizon \( \tau \) for the estimate of joint decision rule \( \sigma_{\tau} \). Thus, the state occupancy is a sufficient statistic for decentralized Markov decision rules since their estimates depend only upon a state occupancy, and no longer on all possible joint observation-histories.

States, belief states, and multi-agent belief states are all sufficient to select directly actions for MDPs, POMDPs, and decentralized POMDPs, respectively. This is mainly because all these statistics summarize the information about the world states from a single agent perspective. The state occupancy, instead, summarizes the information about the world states from the perspective of a team of agents that are constrained to execute their policies independently from each other. In such a setting, joint actions cannot be selected independently, instead, they are selected jointly through decentralized Markov decision rules.

### 4.2 Optimality Criterion

This section presents the optimality criterion based on the policy value functions.

We first define the \( \tau \)-th expected immediate reward function \( r_{\tau}(\cdot; \sigma_{\tau}) : \Delta_{\tau} \rightarrow \mathbb{R} \) that is given by

\[
r_{\tau}(\eta_{\tau}, \sigma_{\tau}) = E_{s \sim \eta_{\tau}} [r(s, \sigma_{\tau}[s])].
\]

This quantity denotes the immediate reward of taking decision rule \( \sigma_{\tau} \) when the system is in state occupancy \( \eta_{\tau} \) at the \( \tau \)-th time step.

Let \( v_{\tau}(\eta_0) \) represent the expected total reward over the decision making horizon if policy \( \pi \) is used and the system is in state occupancy \( \eta_0 \) at the first time step \( \tau = 0 \). For \( \pi \) in the space of decentralized Markov policies, the expected total reward is given by:

\[
v_{\tau}(\eta_0) = \mathbb{E}_{\eta_1, \ldots, \eta_T} \left[ \sum_{\tau=0}^{T-1} r_{\tau}(\eta_{\tau}, \sigma_{\tau}) \mid \eta_0, \pi \right]
\]

We say that a decentralized Markov policy \( \pi^* \) is optimal under the total reward criterion whenever \( v_{\tau^*}(\eta_0) \geq v_{\tau}(\eta_0) \) for all decentralized Markov policies \( \pi \).

Following the Bellman principle of optimality [23], one can separate the problem of finding the optimal policy \( \pi^* \) into simpler subproblems. Each of these subproblems consists of finding policies \( \sigma_{\tau+1} \) that are optimal for all \( \tau = 0, \ldots, T-1 \). To do so, we then define the \( \tau \)-th value function \( v_{\sigma_{\tau+1}^*, \tau+1} : \Delta_{\tau} \rightarrow \mathbb{R} \) under the control of decentralized Markov policy \( \sigma_{\tau+1} \) as follows:

\[
v_{\sigma_{\tau+1}^*, \tau+1}(\eta_{\tau}) = r_{\tau}(\eta_{\tau}, \sigma_{\tau}) + v_{\sigma_{\tau+1}^{*}, \tau+1}(\chi_{\tau+1}(\eta_{\tau}, \sigma_{\tau}))
\]

where quantity \( v_{\sigma_{\tau+1}^*, \tau+1}(\eta_{\tau}) \) denotes the expected sum of rewards attained by starting in state occupancy \( \eta_{\tau} \), taking one joint action according to \( \sigma_{\tau} \), taking the next joint action according to \( \sigma_{\tau+1} \), and so on. We slightly abuse notation and write the \( \tau \)-th value function under the control of an “unknown” decentralized Markov policy \( \sigma_{\tau+1} \) using \( v_{\tau} : \Delta_{\tau} \rightarrow \mathbb{R} \).

We further denote \( \mathcal{V}_{\pi} \) to be the space of bounded value functions at the \( \tau \)-th horizon. For each \( v_{\tau+1} \in \mathcal{V}_{\tau+1} \), and decentralized Markov decision rule \( \sigma_{\tau} \), we define the linear transformation \( \mathcal{L}_{\sigma_{\tau}} : \mathcal{V}_{\tau+1} \rightarrow \mathcal{V}_{\tau} \) by

\[
[\mathcal{L}_{\sigma_{\tau}} v_{\tau+1}](\eta_{\tau}) = r_{\tau}(\eta_{\tau}, \sigma_{\tau}) + v_{\tau+1}(\chi_{\tau+1}(\eta_{\tau}, \sigma_{\tau})).
\]