A. Proof of Proposition 3.2

Proof. We assume that dynamics function F can accurately model the transition function $F(s, \boldsymbol{a}) \approx T(s, \boldsymbol{a})$ and reward function $F(s, \boldsymbol{a}, s') \approx R(s, \boldsymbol{a}, s')$ of the Dec-MDP.

For a policy pair, (π_j, π_k) A MP trajectory consists of a XP trajectory generated from XP joint policy $\boldsymbol{\pi} = (\pi_j, \pi_k)$ starting from starting state $s_0 \sim p(s_0)$ and switching to a SP joint policy $\boldsymbol{\pi} = (\pi_j, \pi_j)$ at timestep h. At $t = h - 1$, the state, s_{h-1} reached by XP joint policy is within $\mathcal{S}_{XP}^{j,k}$. Every subsequent state, $s_{t \geq h}$ is within the set reachable SP states from $\mathcal{S}_{SP}^{j}(s_{xp}), s_{xp} \in \mathcal{S}_{XP}^{j,k}$.

For simulated SP trajectories $\hat{\tau}_F^{\pi_{SP}}(s_{xp})$, the starting states are sampled from reachable XP states $s_{xp} \in \mathcal{S}_{XP}^{j,k}$. The subsequent simulated states are generated via SP joint policy $\boldsymbol{\pi} = (\pi_j, \pi_j)$ and dynamics model F, which accurately models the true environment dynamics. Hence, every simulated state $\hat{s}_{sp} \in \hat{\tau}_F^{\pi_{SP}}(s_{xp})$ is also within the set of reachable SP states from starting state s_{xp}.

□