A Derivation of Policy Optimization Objective

Definitions Consistent with the definition provided in Section 3.2 of the main text, we write $i_{1:m}$ to denote the set of all agents including $i_1, ..., i_m$, $\pi_{1:m} = \pi_{i_1} \times ... \times \pi_{i_m}$, the same applies to $o_{i_{1:m}}$, $s_{i_{1:m}}$, $a_{i_{1:m}}$. A joint policy $\pi_{1:n} : \mathcal{S} \rightarrow \mathcal{A}$ describes a distribution over joint action for each joint state. $\hat{A}^k_t = Q^k_t(s_k, a_k|s_{-k}) - V^k_t(s_k|s_{-k})$ and $r^{\pi_{1:n}}_t(\theta) = \frac{\pi^{\theta}_{1:n}(a^t_{1:n}|s^t_{1:n})}{\pi^{\theta_{old}}_{1:n}(a^t_{1:n}|s^t_{1:n})}$, where $s_{-k} = \{s_1, ..., s_n\} \setminus \{s_k\}$. In addition, if we set $A = a_{1:n}, S = s_{1:n}$, the joint value function and joint Q function of joint policy $\pi_{1:n}$ can be expressed as:

$$V^{\pi_{1:n}}(S) = \sum_t E_{A^t \sim \pi_{1:n}}[\gamma^t R(S^t, A^t)|S^0 = S] \tag{22}$$

$$Q^{\pi_{1:n}}(S, A) = \sum_t E_{A^t \sim \pi_{1:n}}[\gamma^t R(S^t, A^t)|S^0 = S, A^0 = A] \tag{23}$$

In Decision Refinement phase, each agent refine the initial set of paths from the Primordial Planning phase to adjust and avoid specific conflicts. When an agent reaches the nearest point on its planned path, it is awarded a reward of $+r$. Assuming this premise, we posit that the total reward for the agent group at each time step t equals the sum of the rewards for each agent at the corresponding time step:

$$R(S, A) = \sum_{k=1}^{n} r(s_k, a_k|a_{-k}). \tag{24}$$

Based on the aforementioned assumptions and Eqs. (22) and (23), we obtain the following decomposition:

$$V^{\pi_{1:n}}(s_{1:n}) = \sum_{k=1}^{n} V^k(s_k|s_{-k}), \tag{25}$$

$$Q^{\pi_{1:n}}(s_{1:n}, a_{1:n}) = \sum_{k=1}^{n} Q^k(s_k, a_k|s_{-k}), \tag{26}$$

where we define:

$$V^k(s_k|s_{-k}) = E_{a_k \sim \pi_k}[Q^k(s_k, a_k|s_{-k})], \tag{27}$$

$$Q^k(s_k, a_k|s_{-k}) = \sum_t E_{a^t_{-k} \sim \pi_{-k}(.|s_{-k})}[\gamma^t E_{a^t_k \sim \pi_k}[r(s^t_k, a^t_k|a^t_{-k})]|s^0_k = s_k, a^0_k = a_k]. \tag{28}$$

The proof of Eqs.(25) and (26) is provided in Appendix B.1.

Leveraging the previous results, we will consider the multi-agent group as a unified entity and apply the PPO algorithm, yielding the following optimization objective:

$$L_1 = \frac{1}{T} \sum_t min[r^{\pi_{1:n}}_t(\theta) \sum_k \hat{A}^k_t, \text{clip}(r^{\pi_{1:n}}_t(\theta), 1 - \epsilon, 1 + \epsilon) \sum_k \hat{A}^k_t], \tag{29}$$

where $\hat{A}^k_t = Q^k_t(s_k, a_k|s_{-k}) - V^k_t(s_k|s_{-k})$ and $r^{\pi_{1:n}}_t(\theta) = \frac{\pi^{\theta}_{1:n}(a^t_{1:n}|s^t_{1:n})}{\pi^{\theta_{old}}_{1:n}(a^t_{1:n}|s^t_{1:n})}$. Our goal is to decouple the aforementioned collective optimization objective and decompose it into the sum of individual agent optimization objectives, thereby improving the sample utilization rate. We consider the following optimization objective:

$$L_2 = \frac{1}{T} \sum_t \sum_k min[r^{\pi_{1:n}}_t(\theta) \hat{A}^k_t, clip(r^{\pi_{1:n}}_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}^k_t]. \tag{30}$$

We can establish that L_2 serves as a lower bound for L_1:

$$L_1 \geq L_2. \tag{31}$$

Let $\delta = max_{k,t} |\hat{A}^k_t|$, N represents the total number of agents, the estimate for the gap between the two sides of the inequality is as follows:

$$L_1 - L_2 \leq 2\epsilon\delta N. \tag{32}$$

The proof of (31) and (32) is detailed in Appendix B.2. As such, we obtain a decoupled lower bound L_2 for the original centralized optimization objective L_1. However, this lower bound is not applicable to LMAPF, given the assumption of close cooperation among all agents, where the ratio r^k_t of one agent can significantly impact the advantage function allocation of another agent. In the context of solving LMAPF using our PP phase and DR phase framework, the rewards for the agents are largely dispersed rather than aggregated, which would lead to an uneven distribution of the advantage function values if we were to directly use L_2 for optimization. To address this issue, we choose to simply replace the joint agent ratio $r^{\pi_{1:n}}_t$ with the individual agent ratio $r^{\pi_k}_t$. While this approach sacrifices some of the policy-level interaction advantages in the advantage value allocation, the gains exceed the sacrifices due to the dispersed rewards in the PP phase and DR phase solution for LMAPF. Consequently, we have established the following policy optimization objective:

$$L = \frac{1}{T} \sum_t \sum_k min[r^{\pi_k}_t(\theta) \hat{A}^k_t, clip(r^{\pi_k}_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}^k_t]. \tag{33}$$

Owing to the constraints of POMDP and the dispersed characteristics of task rewards, we employ $V^k(s_k|s_{i_{1:m}})$ in lieu of $V^k(s_k|s_{-k})$, $i_{1:m}$ denotes relevant agents within the neighborhood. $\hat{A}^k_t$ is computed based on V^k using the Generalized Advantage Estimation (GAE).

Furthermore, the aforementioned objective function essentially represents an interpolation between MAPPO and IPPO. Specifically, in contrast to MAPPO, which employs global state information for centralized value estimation during training, and IPPO, where each agent independently utilizes its own state information for decentralized value estimation, the proposed objective function leverages both local observations and information from neighboring agents for decision-making and value estimation. This approach is characterized by its locally centralized information aggregation while maintaining decentralized objective generation for individual agents. We hereby term this method as Objective-Decomposed PPO (**OD-PPO**), which specifically designed for the Decision Refinement phase. Notably, we design our neural network based on the information aggregation and decision-making processes of the OD-PPO.

B Mathematical Details

Assumptions

$$R(S,A) = \sum_{k=1}^{n} r(s_k, a_k|a_{-k}) \ .$$

$$V^k(s_k|s_{-k}) = E_{a_k \sim \pi_k}[Q^k(s_k, a_k|s_{-k})] \ .$$

$$Q^k(s_k, a_k|s_{-k}) = \sum_t E_{a^t_{-k} \sim \pi_{-k}(.|s_{-k})}[\gamma^t E_{a^t_k \sim \pi_k}[\\ r(s^t_k, a^t_k|a^t_{-k})]|s^0_k = s_k, a^0_k = a_k] \ .$$

$$L_1 = \frac{1}{T}\sum_t min[r_t^{\pi_{1:n}}(\theta)\sum_k \hat{A}^k_t, \mathrm{clip}(r_t^{\pi_{1:n}}(\theta), 1-\epsilon, \\ 1+\epsilon)\sum_k \hat{A}^k_t] \ .$$

$$L_2 = \frac{1}{T}\sum_t\sum_k min[r_t^{\pi_{1:n}}(\theta)\hat{A}^k_t, clip(r_t^{\pi_{1:n}}(\theta), 1-\epsilon, \\ 1+\epsilon)\hat{A}^k_t] \ .$$

B.1 Multi-Agent Value Decomposition in DR

For any given joint policy $\pi_{1:n}$, joint state $s_{1:n}$ and joint action $a_{1:n}$, we can derive the following decompositions for the joint Q-value and joint V-value:

$$Q^{\pi_{1:n}}(s_{1:n}, a_{1:n}) = \sum_{k=1}^{n} Q^k(s_k, a_k|s_{-k}) \ ,$$

$$\begin{aligned} Proof : & Q^{\pi_{1:n}}(s_{1:n}, a_{1:n}) \\ & = E_{A^t \sim \pi_{1:n}}[\gamma^t R(S^t, A^t)|S^0 = S, A^0 = A] \\ & = \sum_{k=1}^{n}\sum_t E_{A^t \sim \pi_{1:n}}[\gamma^t r(s^t_k, a^t_k|a^t_{-k})|S^0 = s_{1:n}, \\ & \qquad A^0 = a_{1:n}] \\ & = \sum_{k=1}^{n}\sum_t E_{a^t_{-k} \sim \pi_{-k}(.|s_{-k})}[\gamma^t E_{a^t_k \sim \pi_k}[r(s^t_k, a^t_k| \\ & \qquad a^t_{-k})]|s^0_k = s_k, a^0_k = a_k] \\ & = \sum_{k=1}^{n} Q^k(s_k, a_k|s_{-k}) \ . \end{aligned}$$

$$V^{\pi_{1:n}}(s_{1:n}) = \sum_{k=1}^{n} V^k(s_k|s_{-k}) \ ,$$

$$\begin{aligned} Proof : & V^{\pi_{1:n}}(s_{1:n}) \\ & = E_{a_{1:n} \sim \pi_{1:n}}[Q^{\pi_{1:n}}(s_{1:n}, a_{1:n})] \\ & = E_{a_{1:n} \sim \pi_{1:n}}[\sum_{k=1}^{n} Q^k(s_k, a_k|s_{-k})] \\ & = \sum_{k=1}^{n} E_{a_{-k} \sim \pi^{-k}}[E_{a_k \sim \pi^k}[Q^k(s_k, a_k|s_{-k})]] \\ & = \sum_{k=1}^{n} V^k(s_k|s_{-k}) \ . \end{aligned}$$

Lemma 1 When $a > 0$, $b > 0$, and $x_k \in \mathbb{R}$, we obtain the following result:

$$min(a\sum_k x_k, b\sum_k x_k) \geq \sum_k min(ax_k, bx_k) \ ,$$

Proof :Without loss of generality, we assume that:

$$\begin{aligned} & a\sum_k x_k \leq b\sum_k x_k \\ Left & = min(a\sum_k x_k, b\sum_k x_k) = \sum_k ax_k \\ & \geq \sum_k min(ax_k, bx_k) = right \ . \end{aligned}$$

B.2 Upper and lower bound of joint objective

We can prove that L_2 is an lower bound for L_1:

$$L_1 \geq L_2 \ ,$$

Proof : By using Lemma 1, we have:

$$\begin{aligned} L_1 & = \frac{1}{T}\sum_t min[r_t^{\pi_{1:n}}(\theta)\sum_k \hat{A}^k_t, \mathrm{clip}(r_t^{\pi_{1:n}}(\theta), \\ & \qquad 1-\epsilon, 1+\epsilon)\sum_k \hat{A}^k_t] \\ & \geq \frac{1}{T}\sum_t\sum_k min[r_t^{tot}(\theta)\hat{A}^{\pi_k}_t, clip(r_t^{tot}(\theta), \\ & \qquad 1-\epsilon, 1+\epsilon)\hat{A}^{\pi_k}_t] \\ & = L_2 \ . \end{aligned}$$

Let $\delta = max_{k,t}\ |\hat{A}^k_t|$, N represents the total number of agents, the estimate for the gap between the two sides of the inequality is as follows:

$$L_1 - L_2 \leq 2\epsilon\delta N \ ,$$

Proof : Let $a = r_t^{tot}(\theta), b = clip(r_t^{tot}(\theta), 1-\epsilon_N, 1+\epsilon_N)$ and $x_k = A_t^{\pi_k}$, We have:

$$L_1 = min(a\sum_k x_k, b\sum_k x_k) \ ,$$

$$L_2 = \sum_k min(ax_k, bx_k) \ ,$$

Without loss of generality, we assume that:

$$a \geq b > 0 \ ,$$

Owing to the preceding assumptions, it is evident that we have the following result:

when $x_k < 0$, $min(ax_k, bx_k) = ax_k$,

when $x_k \geq 0$, $min(ax_k, bx_k) = bx_k$,

Thus, we have:

$$\sum_k min(ax_k, bx_k) = \sum_{l_1} ax_{l_1} + \sum_{l_2} bx_{l_2} \ ,$$

where $l_1 + l_2 = k, x_{l_1} < 0$ and $x_{l_2} \geq 0$,

We proceed to discuss the cases for $\sum_k x_k$:

For the case where $\sum_k x_k \geq 0$, we have:

$$min(a\sum_k x_k, b\sum_k x_k) = b\sum_k x_k \ .$$

Consequently, the left-hand side of the inequality minus the right-hand side is:

$$\begin{aligned}\Delta &= b\sum_k x_k - a\sum_{l_1} x_{l_1} - b\sum_{l_2} x_{l_2} \\ &= (b-a)\sum_{l_1} x_{l_1} \\ &= (a-b)\sum_{l_1} -x_{l_1} \\ &\leq N|a-b||A|_{max} \ ,\end{aligned}$$

Additionally, by using $|a-b| \leq 2\epsilon$, we have:

$$\Delta \leq 2\epsilon\delta N \ .$$

where $\delta = max_{k,t}\ |\hat{A}_t^k|$, N represents agent number.

Lemma 2 Given finite sets A and B, there exists a bijective function $\mathcal{F} : A \rightarrow B$ with $\mathcal{F}(A) \subseteq B$ and $\mathcal{F}^{-1}(B) \subseteq A$, it follows that $\mathcal{F}(A) = B$ and $\mathcal{F}^{-1}(B) = A$.

Proof : If $\exists b \in B$ satisfy: $\forall a \in A, \mathcal{F}(a) \neq b$.

Let $x = \mathcal{F}^{-1}(b)$.

If $x \in A$ then, $\mathcal{F}(x)$ = b, this is a contradiction.

If $x \notin A$, because of $\mathcal{F}^{-1}(B) \subseteq A$, this is a contradiction.

Thus: $\mathcal{F}(A) = B$, similarly, $\mathcal{F}^{-1}(B) = A$.

B.3 Typical dual transformation

Define $path_{1:n} = \{path_1, ..., path_n\}$ as the trajectories of the agent group $\{1, \ldots, n\}$ within the limited number of time steps. We can prove that symmetric and rotational transformations are Dual Transformations of LMAPF:

Proof : Given a LMAPF problem $U =< G, N, ST, ED >$ we can easily get:

$$\mathcal{F}(U) =< \mathcal{F}(G), N, \mathcal{F}(ST), \mathcal{F}(ED) > \ ,$$

$\forall\ path_{1:n} \in Solution(U)$,

We consider $path_k$, where $k \in \{1, ..., n\}$.

Without loss of generality, we assume that within the limited number of time steps, $path_k$ starts from st_k and passes m_k goals $\{ed_1, \ldots, ed_{m_k}\}$, which means:

$\{ed_1, \ldots, ed_{m_k}\} \subseteq path_k$.

By using $\mathcal{F}(U) =< \mathcal{F}(G), N, \mathcal{F}(ST), \mathcal{F}(ED) >$, we have:

$\mathcal{F}(path_k)$ passing through $\{\mathcal{F}(ed_1), \ldots, \mathcal{F}(ed_{m_k})\}$, which means:

$\{\mathcal{F}(ed_1), \ldots, \mathcal{F}(ed_{m_k})\} \subseteq \mathcal{F}(path_k)$.

Thus: $\bigcup_k \{\mathcal{F}(ed_1), \ldots, \mathcal{F}(ed_{m_k})\} \subseteq \bigcup_k \mathcal{F}(path_k)$.

Then: $\mathcal{F}(Solution(U)) \subseteq Solution(\mathcal{F}(U))$.

Additionally, $\mathcal{F}^{-1}$ is also a rotation/symmetry transformation.

By using Lemma 2, we have:

$\mathcal{F}(Solution(U)) = Solution(\mathcal{F}(U))$.

Lemma 3 Given a LMAPF problem $U =< G, \mathcal{N}, ST, ED >$ and a Dual Transformation $\mathcal{F}$, for any state information s under U, $s^{'}$ under $\mathcal{F}(U)$, we have: $\mathcal{F}(s) = s^{'}$. The proof process is similar to that of B.3.

B.4 Relationship between dual problem and original problem

Given a *Dual Transformation* $\mathcal{F}$, z_p and z_v denoted as the inputs of the policy and value function, we have delineated the respective relationships of optimal policy and optimal value function to $\mathcal{F}(U)$ and U:

$$\mathcal{F}(\pi^*(z_p)) = \pi^*(\mathcal{F}(z_p)),$$

$$V^*(z_v) = V^*(\mathcal{F}(z_v)).$$

In fact, for any state information s under U, we have: $\mathcal{F}(z) = g(F(s))$, where g is the encoder. Thus, for any state information s under U:

$$\mathcal{F}(\pi^*(g_p(s))) = \pi^*(g_p(\mathcal{F}(s))) \ ,$$

$$V^*(g_v(s)) = V^*(g_v(\mathcal{F}(s))) \ .$$

We define $\overline{\pi}^* = \pi^* \circ g_p$, $\overline{V}^* = V^* \circ g_v$, the equations simplify to:

$$\mathcal{F}(\overline{\pi}^*(s)) = \overline{\pi}^*(\mathcal{F}(s)) \ ,$$

$$\overline{V}^*(s) = \overline{V}^*(\mathcal{F}(s)) \ .$$

Proof : Since $\mathcal{F}(Solution(U)) = Solution(\mathcal{F}(U))$,
we have: $\mathcal{F}(\overline{\pi}^*(s)) = \overline{\pi}^*(\mathcal{F}(s))$.
Moreover, $\overline{V}^*(s) = \sum_t E_{a_t \sim \overline{\pi}^*}[r_t(s_t, a_t)|s_0 = s]$.
Let $\mathbf{p}_s = \pi^*(.|s), \mathbf{r}_s = (r(s, a_1), .., r(s, a_{|A|}))$,
$\overline{V}^*(s) = \sum_t \mathbf{p}_{s_t|s_0=s} \cdot \mathbf{r}_{s_t|s_0=s}$.
By using Lemma 3, we have:
$\overline{V}^*(\mathcal{F}(s)) = \sum_t \mathbf{p}_{\mathcal{F}(s_t)|s_0=\mathcal{F}(s)} \cdot \mathbf{r}_{\mathcal{F}(s_t)|s_0=\mathcal{F}(s)}$.
Hence $\mathcal{F}(\overline{\pi}^*(s)) = \overline{\pi}^*(\mathcal{F}(s))$ we have:
$\overline{V}^*(\mathcal{F}(s)) = \sum_t \mathcal{F}(\mathbf{p}_{s_t|s_0=s}) \cdot \mathcal{F}(\mathbf{r}_{s_t|s_0=s})$.
Thus, $\overline{V}^*(s) = \overline{V}^*(\mathcal{F}(s))$.