Algorithm 1: PERLA MAPPO/IPPO

Input: Joint-policy $\boldsymbol{\pi}$, critic parameters $\boldsymbol{\phi}$, policy parameters $\boldsymbol{\theta}$, environment E, number of marginalisation samples K

Output: Optimised joint-policy $\boldsymbol{\pi}$

Instantiate critic $Q_{\boldsymbol{\phi}}$ that takes as input state $\boldsymbol{s}$, action $\boldsymbol{a}_i$ and joint-action $\boldsymbol{a}_{-i}$ `// Equation 2`

Rollout $\boldsymbol{\pi}$ in E to obtain data $D = (\boldsymbol{s}^0, \boldsymbol{a}^1, r^1, \ldots, \boldsymbol{s}^{T-1}, \boldsymbol{a}^T, r^T)$

for $t \leftarrow 0$ **to** T **do**

  for *each agent* i **do**

    Generate K samples of joint-actions $\{\boldsymbol{a}_{-i}^{t(j)} \sim \boldsymbol{\pi}_{-i}(\boldsymbol{\tau}^t)\}_{j=1}^K$

    Compute TD-error $\delta_i = \frac{1}{K}\sum_{j=1}^K Q_\rho(\boldsymbol{s}^{t-1}, \boldsymbol{a}_{-i}^{t-1(j)}, a_i) -$

    $-(\boldsymbol{r} + \gamma \frac{1}{K}\sum_{j=1}^K Q_\rho(\boldsymbol{s}^t, \boldsymbol{a}_{-i}^{t(j)}, a_i^t))$ over sampled joint-actions for each agent `// Equation 2`

    Update critic parameters $\boldsymbol{\rho}$ with δ_i

    Update ith agent's policy parameters $\boldsymbol{\theta}_i$ with advantages given by δ_i, using PPO update

the key benefits of PERLA, namely that it vastly reduces variance of the key estimates used in training. We begin with a result that quantifies the reduction of variance when using $\hat{Q}_i$ instead of Q_i. We defer all proofs to the Appendix.

Theorem 1. *The variance of marginalised Q-function $\tilde{Q}_i$ is smaller than that of the non-marginalised Q-function Q_i for any $i \in \mathcal{N}$, that is to say: Var$(Q_i(s, \mathbf{a})) \geq$ Var$(\tilde{Q}_i(s, a_i))$. Moreover, for the approximation to the marginalised Q-function (c.f. Equation 2) the following relationship holds:*

$$\textit{Var}\left(\hat{Q}_i(s, a_i)\right) = \frac{1}{k}\textit{Var}\left(Q_i(s, \boldsymbol{a}_{-i}, a_i)\right) + \frac{k-1}{k}\textit{Var}\left(\tilde{Q}_i(s, a_i)\right).$$

Therefore for $k = 1$ we get that the approximation has the same variance as the non-marginalised Q-function. However, for any $k > 1$ the approximation to marginalised Q-function has smaller variance that the non-marginalised Q-function. Therefore marginalisation procedure of PERLA can essentially be used as a variance reduction technique. Let us now analyse how this framework can be applied to enhance multi-agent policy gradient algorithm, which is known to suffer from high variance in its original version.

In the policy gradient algorithms, we assume a fully cooperative game which avoids the need to add the agent indices to the state-action and state-value functions since the agents have identical rewards. The goal of each agent is therefore to maximise the expected return from the initial state defined as $\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}_{s_0 \sim p(s_0)}[v(s_0)]$, where $p(s_0)$ is the distribution of initial states and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_N^T)^T$ is the concatenated vector consisting of policy parameters for all agents. The following well-known theorem establishes the gradient of $\mathcal{J}(\boldsymbol{\theta})$ with respect to the policy parameters.

Theorem 2 (MARL Policy Gradient (Zhang et al. 2018))**.**

$$\nabla_{\boldsymbol{\theta}_i}\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t Q(s^t, \boldsymbol{a}_{-i}^t, a_i^t)\nabla_{\boldsymbol{\theta}_i}\log\pi_i(a_i^t|\tau_i^t)\right].$$

Therefore, to calculate the gradient with respect to policy parameters, it is necessary to make use of the state-action-values. In practice, it can estimated by a function approximator, which gives rise to Actor-Critic methods (Konda and Tsitsiklis 1999). In MARL, one can either maintain a centralised critic that provides state-action-value for the i^{th} agent using the knowledge of the other agents' actions or introduce a decentralised critic that does not take actions of others (in its inputs). This gives rise to the centralised training decentralised execution (CT-DE) $\boldsymbol{g}_i^{\mathrm{C}}$ and decentralised $\boldsymbol{g}_i^{\mathrm{D}}$ gradient estimators respectively, as defined below.

$$\boldsymbol{g}_i^{\mathrm{C}} := \sum_{t=0}^{\infty}\gamma^t Q(s^t, \boldsymbol{a}_{-i}^t, a_i^t)\nabla_{\boldsymbol{\theta}_i}\log\pi_i(a_i^t|\tau_i^t), \quad (3)$$

$$\boldsymbol{g}_i^{\mathrm{D}} := \sum_{t=0}^{\infty}\gamma^t \tilde{Q}(s^t, a_i^t)\nabla_{\boldsymbol{\theta}_i}\log\pi_i(a_i^t|\tau_i^t), \quad (4)$$

where $Q(s_t, a_{-i}^t, a_i^t)$ and $\tilde{Q}(s^t, a_i^t)$ are the CT-DE and decentralised critics respectively. In PERLA Actor-Critic algorithms (such as in Algorithm 1) we utilise a third estimator - the PERLA estimator given below:

$$\boldsymbol{g}_i^{\mathrm{P}} := \sum_{t=0}^{\infty}\gamma^t \hat{Q}(s^t, a_i^t)\nabla_{\boldsymbol{\theta}_i}\log\pi_i(a_i^t|\tau_i^t),$$

where $\hat{Q}(s^t, a_i^t)$ is the Monte-Carlo approximation of $\tilde{Q}(s^t, a_i^t)$. Note that in this approach we maintain a centralised critic $Q(s^t, a_{-i}^t, \boldsymbol{a}_i^t)$, therefore the approximation to the marginalised Q-function is obtained using $\hat{Q}(s^t, a_i^t) = \frac{1}{k}\sum_{j=1}^k Q(s^t, a_i^t, \boldsymbol{a}_{-i}^{t(j)})$, where $\boldsymbol{a}_{-i}^{t(j)} \sim \pi_{-i}(\boldsymbol{a}_{-i}^t|\tau_{-i}^t)$. The PERLA estimator is equal in expectation to the CT-DE estimator as stated by the following Theorem.

Theorem 3. *Given the same (possibly imperfect) critic, the estimators $\boldsymbol{g}_i^C$ and $\boldsymbol{g}_i^P$ have the same expectation, that is:*

$$\mathbb{E}[\boldsymbol{g}_i^C] = \mathbb{E}[\boldsymbol{g}_i^P].$$

Therefore, whenever the critic provides the true Q-value, PERLA estimator is an unbiased estimate of the policy gradient (which follows from Theorem 2). However, although the CT-DE and PERLA estimators have the same expectations, the PERLA estimator enjoys significantly lower variance. Similar to the analysis in (Kuba et al. 2021), let us study the excess variance those two estimators have over the decentralised estimator. Let B_i be the upper bound on the gradient norm of ith agent, i.e. $B_i = \sup_{s,\boldsymbol{a}}||\nabla_{\theta^i}\log\pi_i(a_i|s)||$ and C be the upper bound on the Q-function, i.e. $C = \sup_{s,\boldsymbol{a}} Q(s, \boldsymbol{a})$.

We can now present two theorems showing the effectiveness of PERLA estimator for policy gradients.

Theorem 4. *Given true Q-values, the difference in variances between the decentralised and PERLA estimators admits the following bound:*

$$\textit{Var}(\boldsymbol{g}_i^P) - \textit{Var}(\boldsymbol{g}_i^D) \leq \frac{1}{k}\frac{B_i^2 C^2}{1-\gamma^2}.$$