Federated Reinforcement Learning with Environment Heterogeneity

Hao Jin
Peking University

Yang Peng
Peking University

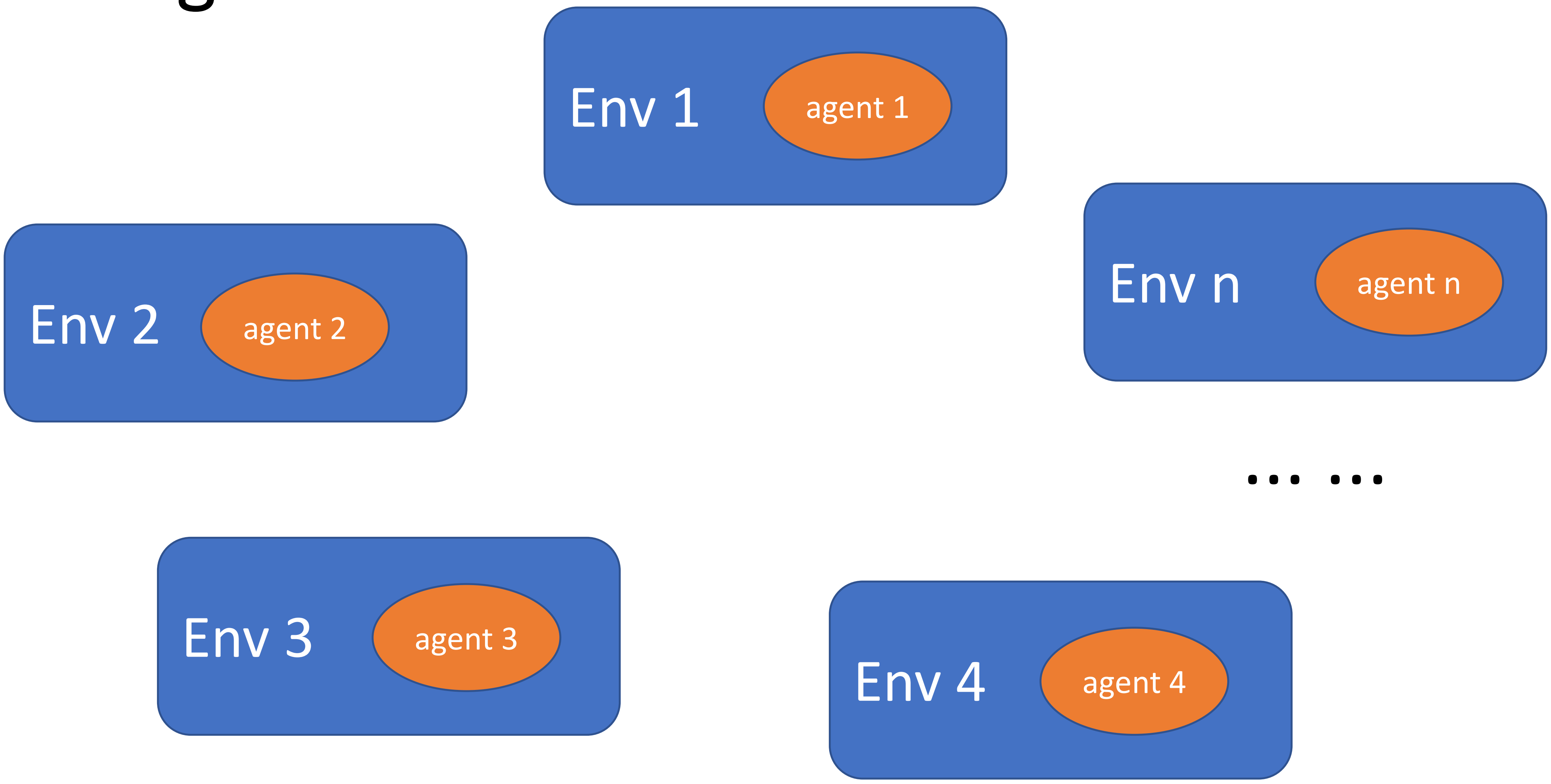
Wenhao Yang
Peking University

Shusen Wang
Xiaohongshu Inc.

Zhihua Zhang
Peking University

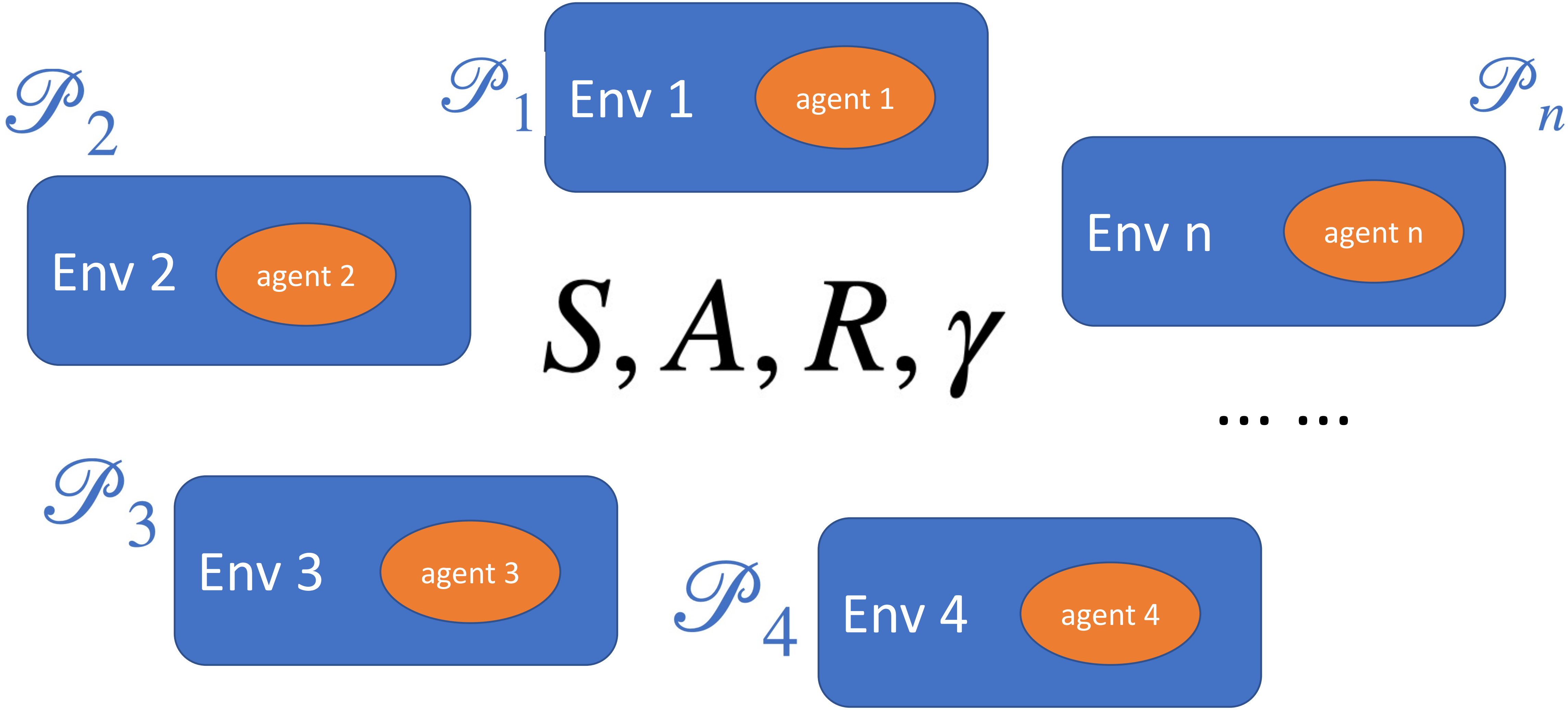
AISTATS 2022

Setting

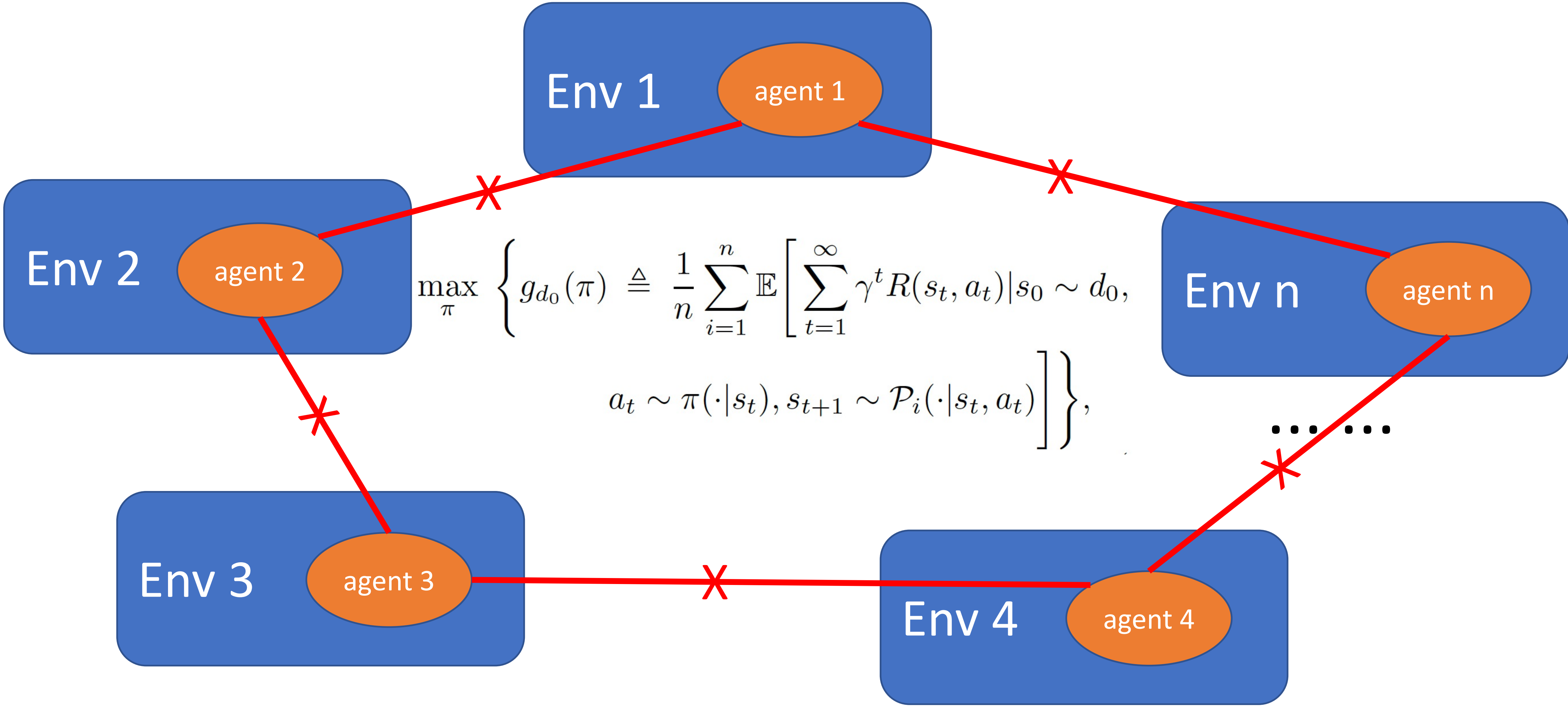


Env 1
agent 1
Env 2
agent 2
Env n
agent n
… …
Env 3
agent 3
Env 4
agent 4

Environment Heterogeneity



Federated Setting



$$\max_{\pi} \left\{ g_{d_0}(\pi) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, a_t) | s_0 \sim d_0, \right.\right.$$
$$\left.\left. a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}_i(\cdot|s_t, a_t) \right] \right\},$$

QAvg & PAvg

Methodology: Periodic communication of policies.

QAvg:

$$Q_{t+1}^k(s,a) \leftarrow (1-\eta_t)\cdot Q_t^k(s,a) + \eta_t \cdot \Big[R(s,a) + \gamma \sum_{s'} \mathcal{P}_k(s'|s,a) \max_{a'\in\mathcal{A}} Q_t^k(s',a')\Big].$$

$$\bar{Q}_t(s,a) \leftarrow \frac{1}{n}\sum_{i=1}^{n} Q_t^i(s,a),\ \forall\, s,a;$$

$$Q_t^i(s,a) \leftarrow \bar{Q}_t(s,a),\ \forall\, s,a,k.$$

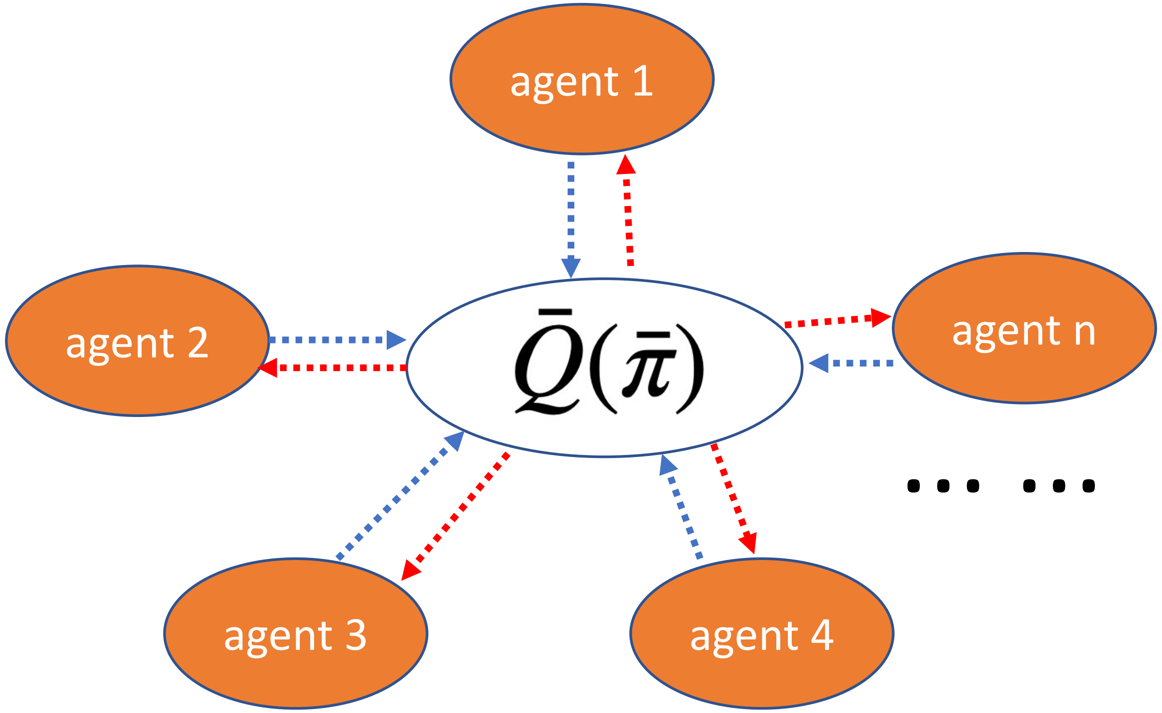
PAvg:

$$\tilde{\pi}_{t+1}^k(a|s) \leftarrow \pi_t^k(a|s) + \frac{\partial g_{d_0,k}(\pi_t^k)}{\partial \pi(a|s)},\ \forall\, s,a,k;$$

$$\pi_{t+1}^k(\cdot|s) \leftarrow \mathop{\mathrm{Proj}}_{\Delta(\mathcal{A})}(\tilde{\pi}_{t+1}^k(\cdot|s)),\ \forall\, s,a,k.$$

$$\bar{\pi}_t(a|s) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \pi_t^i(a|s),\ \forall\, s,a;$$

$$\pi_t^i(a|s) \leftarrow \bar{\pi}_t(a|s),\ \forall\, s,a,k.$$



Theoretical analysis

Environment Heterogeneity: scalars to quantify such heterogeneity.

$$\kappa_1 \triangleq \max_{s,\pi} \sum_{s'} \sum_{i=1}^{n} \left| \mathcal{P}_i^{\pi}(s'|s) - \frac{1}{n} \sum_{j=1}^{n} \mathcal{P}_j^{\pi}(s'|s) \right|,$$

$$\kappa_2 \triangleq \max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \left\| \nabla_{\pi} g_{d_0,i}(\pi) - \frac{1}{n} \sum_{j=1}^{n} \nabla_{\pi} g_{d_0,j}(\pi) \right\|_2,$$

Theoretical analysis (QAvg)

Imaginary Environment: same S, A, R, γ

but with an averaged dynamic $\bar{\mathcal{P}}(s'|s,a) = \frac{1}{n}\sum_{k=1}^{n} \mathcal{P}_k(s'|s,a), \ \forall s, s' \in \mathcal{S}, \ \forall a \in \mathcal{A}.$

Convergent Results:

- Q table of optimal policy in imaginary environment.
- Optimality gap controlled by environment heterogeneity κ_1

Theoretical analysis (PAvg)

Convergent Results:

- Performance gap controlled by environment heterogeneity κ_2
- Performance gap affected by the communication period length ***E***.

$$\max_{\pi} \left\{ g_{d_0}(\pi) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, a_t) | s_0 \sim d_0, \right.\right.$$
$$\left.\left. a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}_i(\cdot|s_t, a_t) \right] \right\},$$

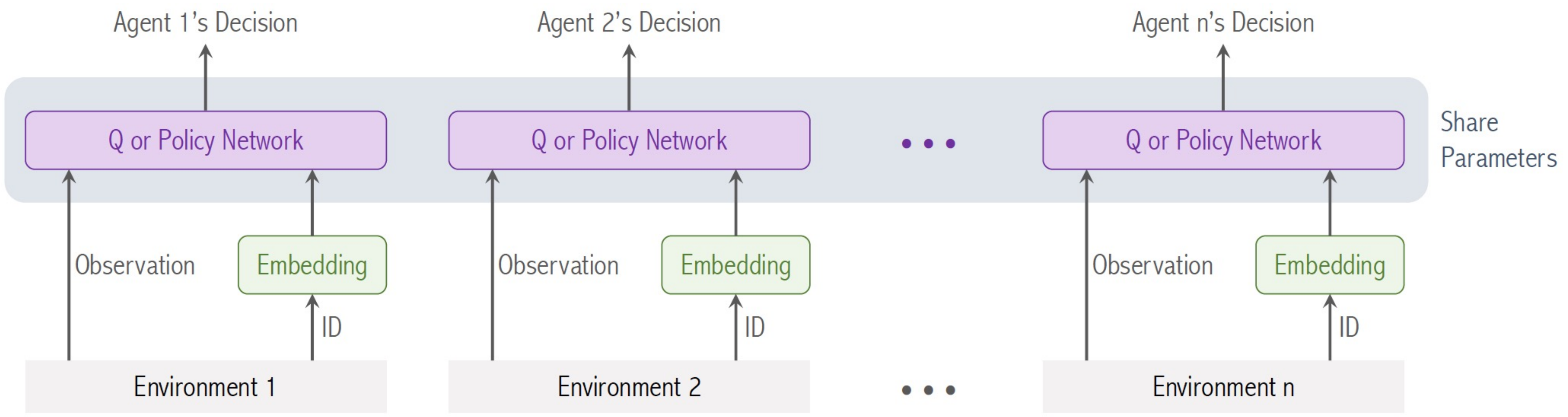
Personalization heuristic

Sub-optimality issues:

- The federated learned policy is sub-optimal in any local env.

Personalization heuristic:

- Every agent learns a private embedding to characterize its local env.



Numerical Experiments

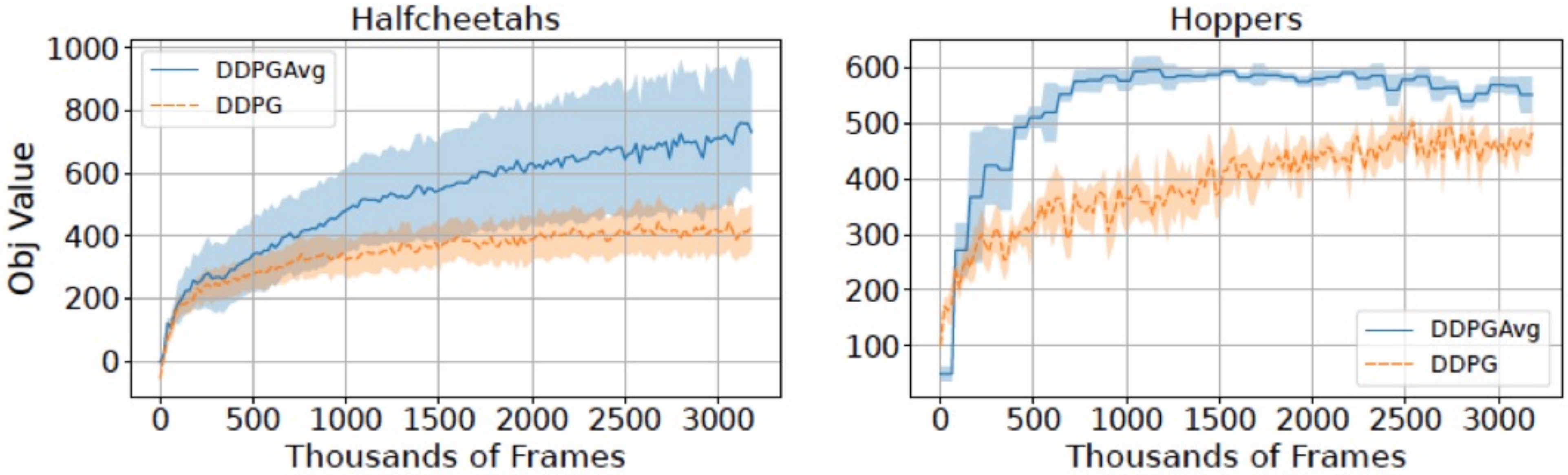
Justification of QAvg and PAvg:

- Larger environment heterogeneity leads to greater optimality gap.
- Different numbers of iterations between communication affect convergence of PAvg.

Numerical Experiments

Deep extensions on DQNAvg and DDPGAvg:

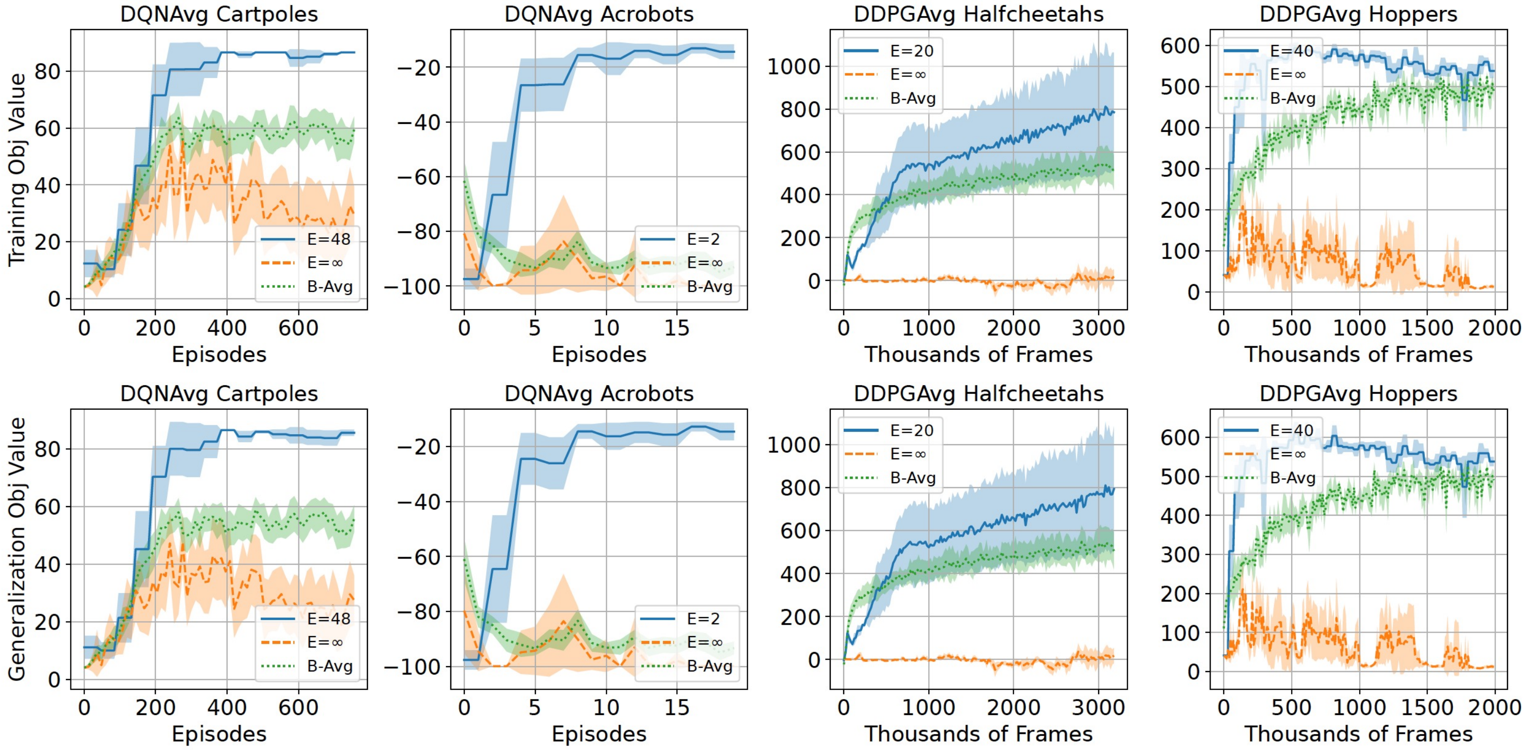
- Faster learning with policy communication.



Numerical Experiments

Deep extensions on DQNAvg and DDPGAvg:

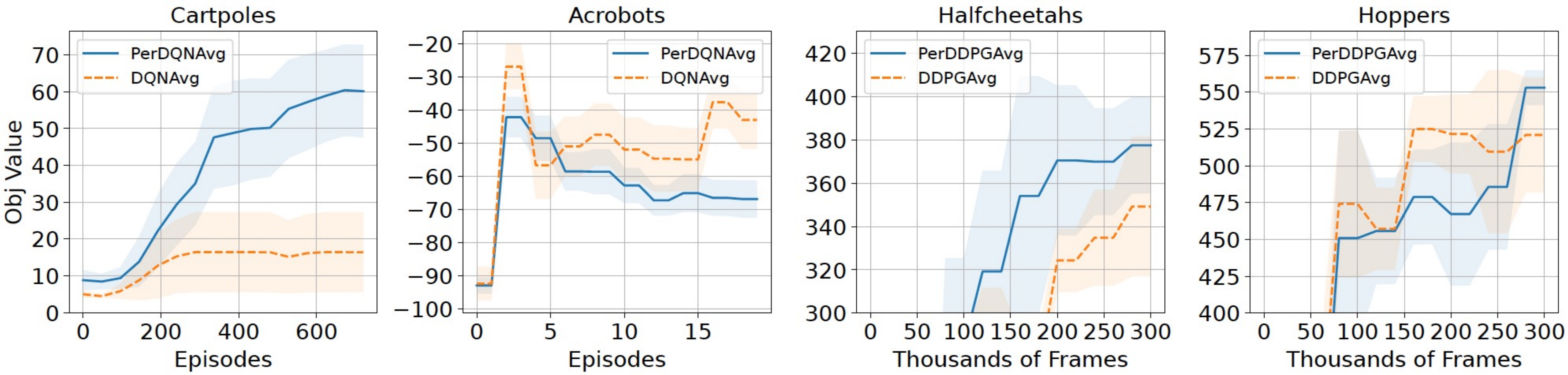
- Faster learning with policy communication.
- Better performance on similar but unseen environment.



Numerical Experiments

Performance of personalization heuristic:

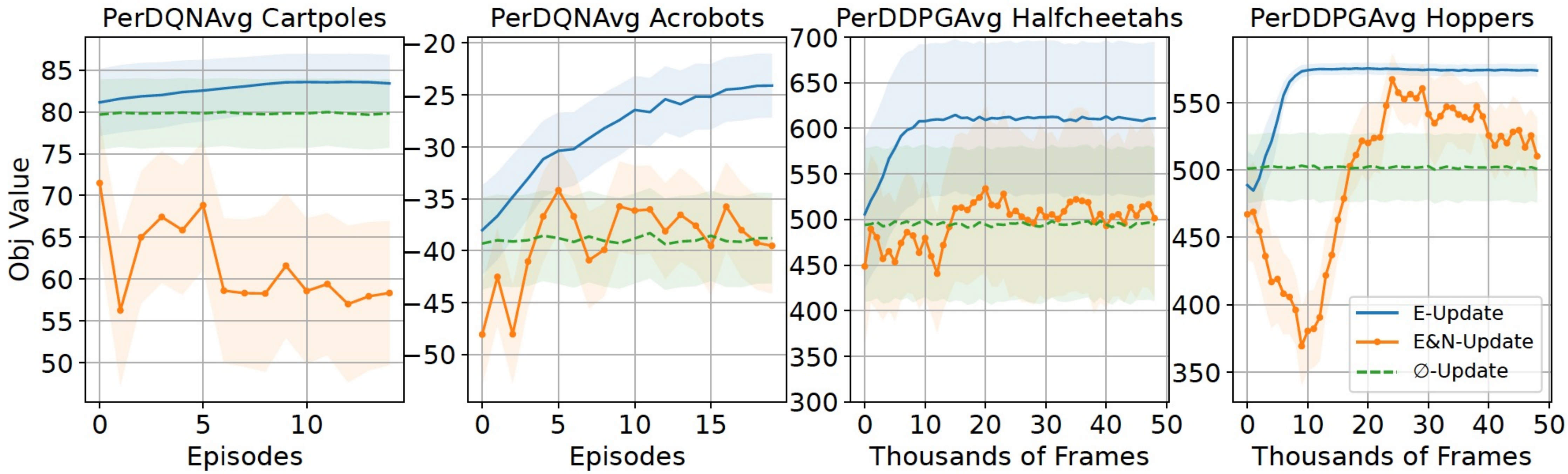
- Decrease the optimality gap in each environment.



Numerical Experiments

Performance of personalization heuristic:

- Decrease the optimality gap in each environment.
- Enable faster generalization to an unseen but similar environment.



Thank you!