Safe and Robust Experience Sharing for Deterministic Policy Gradient Algorithms

Baturay Saglam, Dogan C. Cicek, Furkan B. Mutlu and Suleyman S. Kozat

Department of Electrical and Electronics Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey

Problem Description

1. Off-policy deep reinforcement learning requires large interactions with the environment to obtain optimal policies. Therefore, agents should obtain optimal policies even when the experience replay buffer is very limited [1].
2. Experience sharing is an effective alternative for this problem. However, learning from other agents' experiences may lead to the *extrapolation error* [1].
3. Importance sampling can overcome the extrapolation error using action probability estimates. However, it is not a valid option for deterministic policies by nature.

Contributions

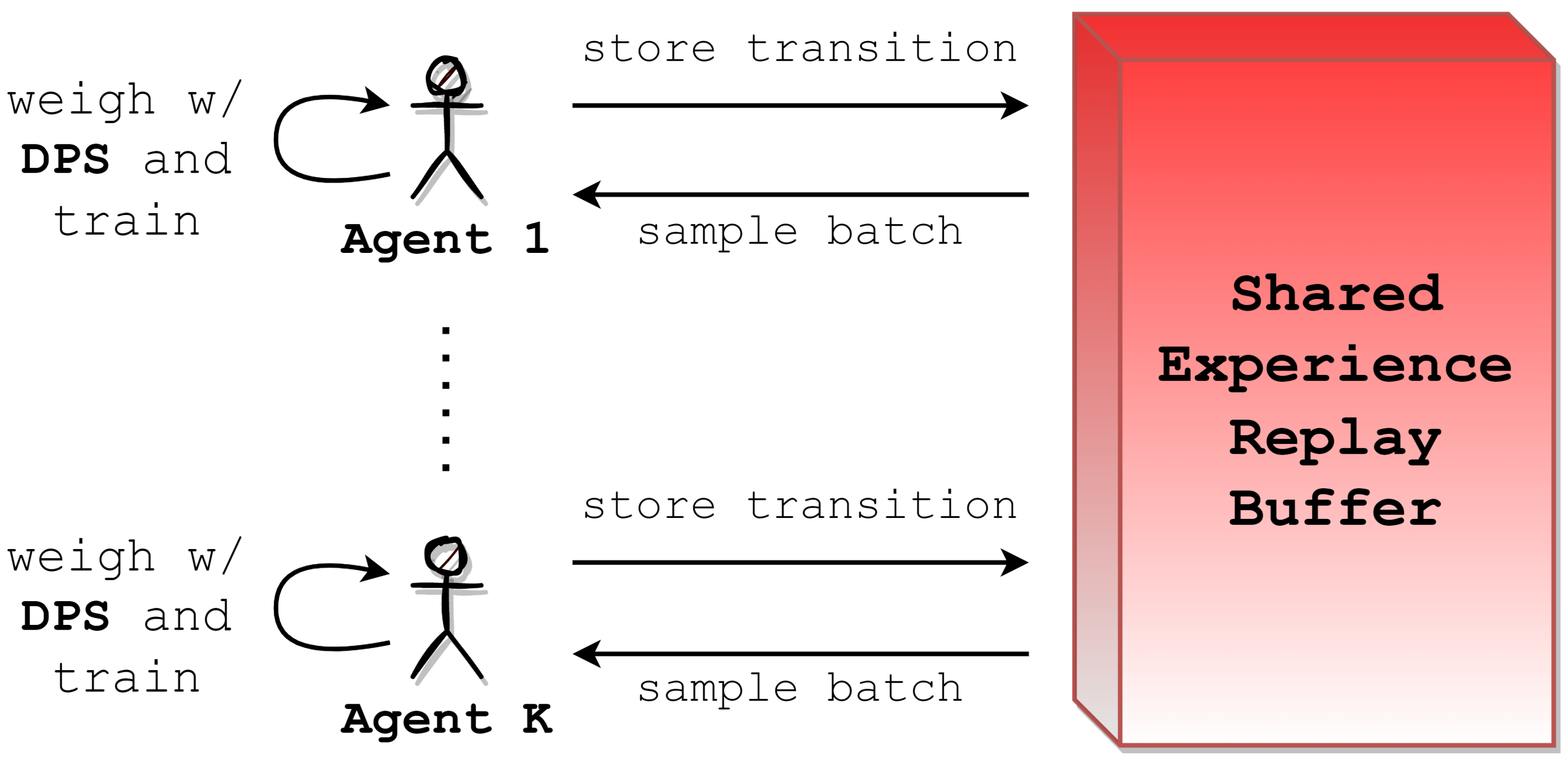
1. We introduce a policy similarity measurement for deterministic policies without any action probability estimates.
2. Using the presented measure, we consitute an actor-critic architecture that involves multiple explorer agents that share experience with each other.
3. Due to the exploration across different copies of the same environment, our algorithm enables a diverse exploration and offers a faster convergence to higher rewards.

Notation

1. Each agents samples a batch of transitions from the shared replay buffer: $(\boldsymbol{S}^{|\mathcal{B}|\times m}, \boldsymbol{A}^{|\mathcal{B}|\times n}, \boldsymbol{R}^{|\mathcal{B}|\times 1}, \boldsymbol{S}'^{|\mathcal{B}|\times m}) \sim \mathcal{B}$
2. Each batch can contain external (collected by other agents) and internal (collected by the agent in interest) transitions: $\mathcal{B}_I \cup \mathcal{B}_E = \mathcal{B}$ and $\mathcal{B}_I \cap \mathcal{B}_E = \emptyset$.

Algorithm

1. Compute the current action decisions on the external states: $\hat{\boldsymbol{A}}_E^{|\mathcal{B}_E|\times n} = \pi_\phi(\boldsymbol{S}_E^{|\mathcal{B}_E|\times m})$
2. Obtain the action difference batch: $\dot{\boldsymbol{A}}^{|\mathcal{B}_E|\times n} \coloneqq \boldsymbol{A}_E^{|\mathcal{B}_E|\times n} - \hat{\boldsymbol{A}}_E^{|\mathcal{B}_E|\times n}$
3. Compute the mean of the the multivariate Gaussian: $\dot{\boldsymbol{\mu}}^{n\times 1} = \frac{1}{|\mathcal{B}_E|}\sum_{i=1}^{|\mathcal{B}_E|}(\dot{\boldsymbol{A}}_i^{|\mathcal{B}_E|\times n})^\top$
4. Compute the covariance matrix of the multivariate Gaussian: $\dot{\boldsymbol{\Sigma}}^{n\times n} = \frac{1}{|\mathcal{B}_E|-1}\sum_{i=1}^{|\mathcal{B}_E|}\boldsymbol{a}_i^{n\times 1}(\boldsymbol{a}_i^{n\times 1})^\top$
5. Compute the dissimilarity metric: $\rho = \mathrm{JSD}(\mathcal{N}(\dot{\boldsymbol{\mu}}^{n\times 1}, \dot{\boldsymbol{\Sigma}}^{n\times n}) \parallel \mathcal{N}(\boldsymbol{0}^{n\times 1}, \sigma \boldsymbol{I}^{n\times n}))$
6. Convert the dissimilarity to the similarity to construct the **Deterministic Policy Similarity (DPS)** weights: $\boldsymbol{\lambda}^{|\mathcal{B}|_E\times 1} = [e^{-\rho},\ e^{-\rho},\ \ldots,\ e^{-\rho}]^\top$
7. Weigh the external transitions by $\boldsymbol{\lambda}^{|\mathcal{B}|_E\times 1}$
8. Train the policy and value networks



weigh w/ **DPS** and train

Agent 1

store transition

sample batch

weigh w/ **DPS** and train

Agent K

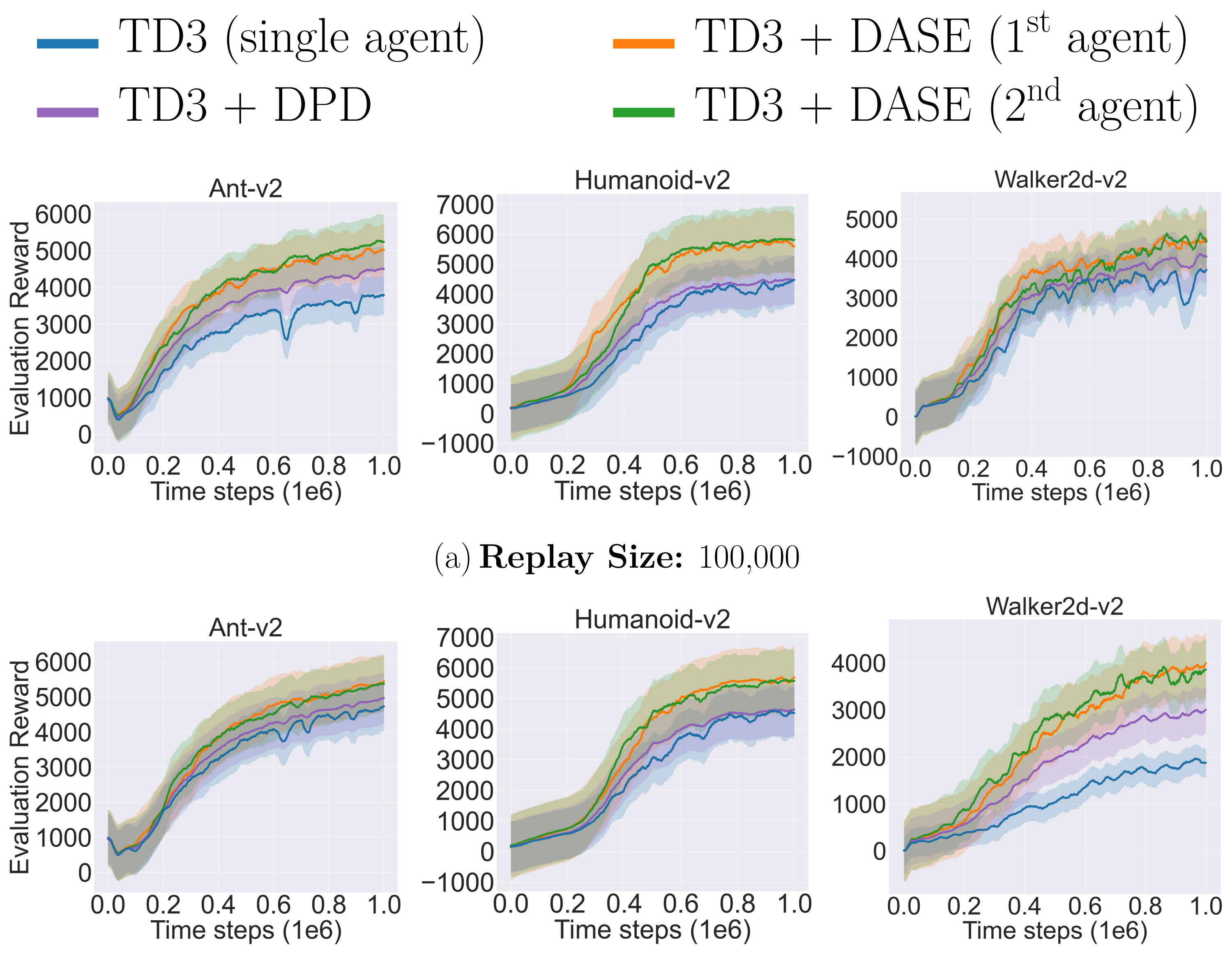
store transition

sample batch

Shared Experience Replay Buffer

Figure 1:Deterministic Actor-Critic with Shared Experience (DASE) architecture

Results



(a) **Replay Size:** 100,000

(b) **Replay Size:** 1,000,000

Figure 2:Learning curves for the set of OpenAI Gym continuous control tasks when replay size is 1 million and 100,000. The shaded region represents half a standard deviation of the average evaluation return over 10 random seeds. Twin Delayed Deep Deterministic Policy Gradient (TD3) [2] is used as the baseline actor-critic algorithm and the method is compared with Dual Policy Distillation (DPD) [3].

Conclusion

We introduce a novel multi-agent actor-critic architecture that enables robust parallel learning for deterministic policies. By safely sharing experience across multiple agents, the presented architecture diversifies the explored state space and overcomes the extrapolation error. The experimental results show that it can obtain state-of-the-result when the replay memory is strictly limited, for which the competing approaches are stuck at suboptimal policies.

References

[1] Scott Fujimoto, David Meger, and Doina Precup.
Off-policy deep reinforcement learning without exploration, 2018.

[2] Scott Fujimoto, Herke van Hoof, and David Meger.
Addressing function approximation error in actor-critic methods, 2018.

[3] Kwei-Herng Lai, Daochen Zha, Yuening Li, and Xia Hu.
Dual policy distillation, 2020.