Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

옥찬호
utilForever@gmail.com

Prerequisites

- Markov Decision Process
- Policy Iteration
- Policy Gradient
- Actor-Critic
- Entropy
- KL Divergence
- PPO (Proximal Policy Optimization)
- GAN (Generative Adversarial Network)

Introduction

- Generalization and exploration in RL still represent key challenges that leave most current methods ineffective.
  - First, a battery of recent studies (Farebrother et al., 2018; Zhang et al., 2018a; Song et al., 2020; Cobbe et al., 2020) indicates that current RL methods fail to generalize correctly even when agents have been trained in a diverse set of environments.
  - Second, exploration has been extensively studied in RL; however, most hard-exploration problems use the same environment for training and evaluation.

Introduction

- Hence, since a well-designed exploration strategy should maximize the information received from a trajectory about an environment, **the exploration capabilities may not be appropriately assessed if that information is memorized**.

Introduction

- In this work, we propose Adversarially Guided Actor-Critic (AGAC), which reconsiders the actor-critic framework by introducing a third protagonist: **the adversary**.
  - Its role is to predict the actor's actions correctly. Meanwhile, the actor must not only find the optimal actions to maximize the sum of expected returns, but also counteract the predictions of the adversary.
  - This formulation is lightly inspired by adversarial methods, specifically generative adversarial networks (GANs) (Goodfellow et al., 2014).

Introduction

- Such a link between GANs and actor-critic methods has been formalized by Pfau & Vinyals (2016); however, in the context of a third protagonist, we draw a different analogy.
  - <u>The adversary</u> can be interpreted as playing <u>the role of a discriminator that must predict the actions of the actor</u>, and <u>the actor</u> can be considered as playing <u>the role of a generator that behaves to deceive the predictions of the adversary</u>.
  - This approach has the advantage, as with GANs, that the optimization procedure <u>generates a diversity of meaningful data</u>, corresponding to sequences of actions in AGAC.

Introduction

- The contributions of this work are as follow:
  - (i) we propose a novel actor-critic formulation inspired from adversarial learning (AGAC)
  - (ii) we analyze empirically AGAC on key reinforcement learning aspects such as diversity, exploration and stability
  - (iii) we demonstrate significant gains in performance on several sparse-reward hard-exploration tasks including procedurally-generated tasks

Backgrounds and Notations

- Markov Decision Process (MDP)

$$M = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma\}$$

  - $\mathcal{S}$: The state space
  - $\mathcal{A}$: The action space
  - $\mathcal{P}$: The transition kernel
  - R : The bounded reward function
  - $\gamma \in [0, 1)$: The discount factor

Backgrounds and Notations

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Let π denote a stochastic policy mapping states to distributions over actions. We place ourselves in the infinite-horizon setting.
  - i.e., we seek a policy that optimizes $J(\pi) = [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$.
  - The value of a state is the quantity $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$ and the value of a state-action pair $Q^\pi(s, a)$ of performing action a in state s and then following policy π is defined as: $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$.
  - The advantage function, which quantifies how an action a is better than the average action in state s, is $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$.
  - Finally, the entropy $\mathcal{H}^\pi$ of a policy is calculated as: $\mathcal{H}^\pi(s) = \mathbb{E}_{\pi(\cdot|s)}[-\log \pi(\cdot\,|s)]$.

Backgrounds and Notations

- An actor-critic algorithm is composed of two main components: a policy and a value predictor.
  - In deep RL, both the policy and the value function are obtained via parametric estimators: we denote θ and ϕ their respective parameters.
  - The policy is updated via policy gradient, while the value is usually updated via temporal difference or Monte Carlo rollouts.

Backgrounds and Notations

- For a sequence of transitions $\{s_t, a_t, r_t, s_{t+1}\}_{t\in[0,N]}$, we use the following policy gradient loss (including the commonly used entropic penalty):

$$\mathcal{L}_{PG} = -\frac{1}{N}\sum_{t'=t}^{t+N}\Big(\underbrace{A_{t'}}_{\text{Critic}}\underbrace{\log\pi\,(a_{t'}|s_{t'},\theta)}_{\text{Actor}} + \underbrace{\alpha\mathcal{H}^{\pi}(s_{t'},\theta)}_{\text{Entropy}}\Big)$$

where α is the entropy coefficient and A_t is the generalized advantage estimator (Schulman et al., 2016) defined as:

$$A_t = \sum\nolimits_{t'=t}^{t+N}(\gamma\lambda)^{t'-t}(r_{t'} + \gamma V_{\phi_{\text{old}}}(s_{t'+1}) - V_{\phi_{\text{old}}}(s_{t'}))$$

with λ a fixed hyperparameter and $V_{\phi_{\text{old}}}$ the value function estimator at the previous optimization iteration.

Backgrounds and Notations

- To estimate the value function, we solve the non-linear regression problem

  $\text{minimize}_{\phi} \sum_{t'=t}^{t+N} \left(V_{\phi}(s_{t'}) - \hat{V}_{t'}\right)^2$ where $\hat{V}_{t'} = A_t + V_{\phi_{old}}(s_{t'})$.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- To foster diversified behavior in its trajectories, AGAC introduces a third protagonist to the actor-critic framework: **the adversary**.
  - The role of the adversary is to accurately predict the actor's actions, by **minimizing** the discrepancy between its action distribution π_{adv} and the distribution induced by the policy π.
  - Meanwhile, in addition to finding the optimal actions to **maximize** the sum of expected returns, the actor must also counteract the adversary's predictions by **maximizing** the discrepancy between π and π_{adv}.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

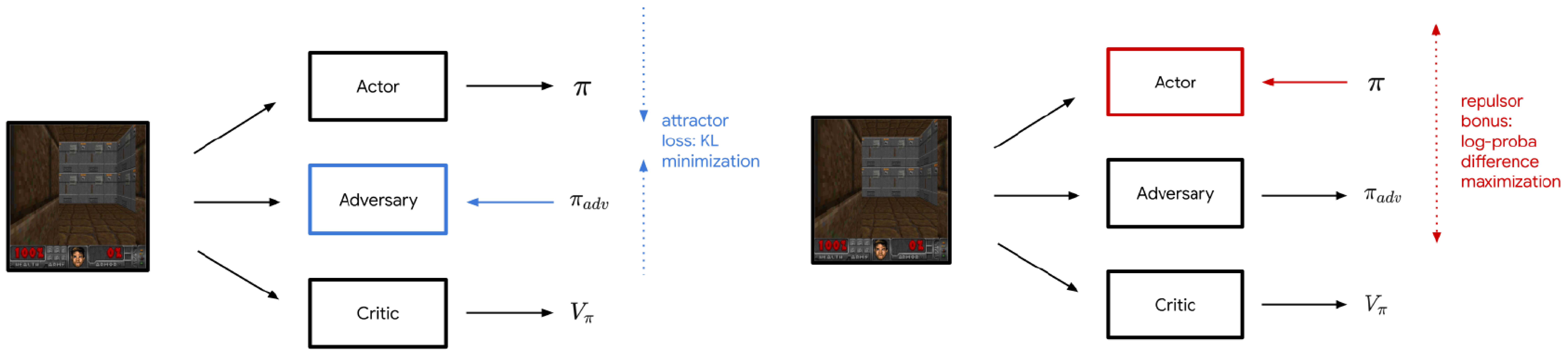


Figure 10: A simple schematic illustration of AGAC. **Left:** the adversary minimizes the KL-divergence with respect to the action probability distribution of the actor. **Right:** the actor receives a bonus when counteracting the predictions of the adversary.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- This discrepancy, used as a form of exploration bonus, is defined as **the difference of action log-probabilities**, whose expectation is the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}\left(\pi(\cdot\,|s)\,\|\,\pi_{\mathrm{adv}}(\cdot\,|s)\right) = \mathbb{E}_{\pi(\cdot|s)}[\log \pi(\cdot\,|s) - \log \pi_{\mathrm{adv}}(\cdot\,|s)]$$

- Formally, for each state-action pair (s_t, a_t) in a trajectory, an action-dependent bonus $\log \pi\,(a_t|s_t) - \log \pi_{\mathrm{adv}}\,(a_t|s_t)$ is added to the advantage.

Adversarially Guided AC

Adversarially Guided Actor-Critic, Y. Flet-Berliac et al, 2021

- The value target of the critic is modified to include the action-independent equivalent, which is the KL-divergence $D_{\mathrm{KL}}\left(\pi(\cdot\,|s)\|\pi_{\mathrm{adv}}(\cdot\,|s)\right)$.
- In addition to the parameters θ (resp. θ_{old} the parameter of the policy at the previous iteration) and ϕ defined above (resp. ϕ_{old} that of the critic), we denote ψ (resp. ψ_{old}) that of the adversary.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- AGAC minimizes the following loss:

$$\mathcal{L}_{\text{AGAC}} = \mathcal{L}_{\text{PG}} + \beta_V \mathcal{L}_V + \beta_{\text{adv}} \mathcal{L}_{\text{adv}}$$

  - In the new objective $\mathcal{L}_{\text{PG}} = \frac{1}{N}\sum_{t=0}^{N}\left(A_t^{\text{AGAC}} \log \pi\,(a_t|s_t,\theta) + \alpha \mathcal{H}^{\pi}(s_t,\theta)\right)$,

    AGAC modifies A_t as:

$$A_t^{\text{AGAC}} = A_t + c\big(\log \pi\,(a_t|s_t,\theta_{\text{old}}) - \log \pi_{\text{adv}}\,(a_t|s_t,\psi_{\text{old}})\big)$$

    with c is a varying hyperparameter that controls the dependence on the action log-probability difference.

    To encourage exploration without preventing asymptotic stability, c is linearly annealed during the course of training.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- $\mathcal{L}_V$ is the objective function of the critic defined as:

$$\mathcal{L}_V = \frac{1}{N}\sum_{t=0}^{N}\left(V_\phi(s_t) - \left(\hat{V}_t + cD_{\mathrm{KL}}\left(\pi(\cdot\,|s_t,\theta_{\mathrm{old}})\|\pi_{\mathrm{adv}}(\cdot\,|s_t,\psi_{\mathrm{old}})\right)\right)\right)^2$$

- Finally, $\mathcal{L}_{\mathrm{adv}}$ is the objective function of the adversary:

$$\mathcal{L}_{\mathrm{adv}} = \frac{1}{N}\sum_{t=0}^{N} D_{\mathrm{KL}}\left(\pi(\cdot\,|s_t,\theta_{\mathrm{old}})\|\pi_{\mathrm{adv}}(\cdot\,|s_t,\psi)\right)$$

- They are the three equations that our method modifies in the traditional actor-critic framework. The terms β_V and β_{adv} are fixed hyperparameters.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Under the proposed actor-critic formulation, the probability of sampling an action is increased if the modified advantage is positive, i.e.
  - (i) the corresponding return is larger than the predicted value
  - (ii) the action log-probability difference is large
- More precisely, our method favors transitions whose actions were less accurately predicted than the average action, i.e.

$$\log \pi(a|s) - \log \pi_{\mathrm{adv}}(a|s) \geq D_{\mathrm{KL}}\left(\pi(\cdot\,|s) \middle\| \pi_{\mathrm{adv}}(\cdot\,|s)\right).$$

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- This is particularly visible for $\lambda \to 1$, in which case the generalized advantage is $A_t = G_t - V_{\phi_{\text{old}}}(s_t)$, resulting in the appearance of both aforementioned mirrored terms in the modified advantage:

$$A_t^{\text{AGAC}} = G_t - \hat{V}_t^{\phi_{\text{old}}} + c\left(\log \pi\,(a_t|s_t) - \log \pi_{\text{adv}}\,(a_t|s_t) - \widehat{D}_{\text{KL}}^{\phi_{\text{old}}}\left(\pi(\cdot\,|s)\|\pi_{\text{adv}}(\cdot\,|s)\right)\right)$$

with G_t the observed return, $\hat{V}_t^{\phi_{\text{old}}}$ the estimated return and $\widehat{D}_{\text{KL}}^{\phi_{\text{old}}}\left(\pi(\cdot\,|s)\|\pi_{\text{adv}}(\cdot\,|s)\right)$ the estimated KL-divergence.

Adversarially Guided AC

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- To avoid instability, in practice the adversary is a separate estimator, updated with a smaller learning rate than the actor. This way, it represents a delayed and more steady version of the actor's policy, which prevents the agent from having to constantly adapt or focus solely on fooling the adversary.

Building Motivation

- We provide an interpretation of AGAC by studying the dynamics of attraction and repulsion between the actor and the adversary.
- To simplify, we study the equivalent of AGAC in a policy iteration (PI) scheme. PI being the dynamic programming scheme underlying the standard actor-critic, we have reasons to think that some of our findings translate to the original AGAC algorithm.

Building Motivation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- In PI, the quantity of interest is the action-value, which AGAC would modify as:

$$Q_{\pi_k}^{\mathrm{AGAC}} = Q_{\pi_k} + c(\log \pi_k - \log \pi_{\mathrm{adv}})$$

with π_k the policy at iteration k.

- Incorporating the entropic penalty, the new policy π_{k+1} verifies:

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}\, \mathcal{J}_{\mathrm{PI}}(\pi) = \underset{\pi}{\operatorname{argmax}} \mathbb{E}_s \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_{\pi_k}^{\mathrm{AGAC}}(s,a) - \alpha \log \pi(a|s)\right]$$

Building Motivation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- We can rewrite this objective:

$$
\begin{aligned}
&\mathcal{J}_{\mathrm{PI}}(\pi) \\
&= \mathbb{E}_s \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_{\pi_k}^{\mathrm{AGAC}}(s,a) - \alpha \log \pi(a|s)\right] \\
&= \mathbb{E}_s \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_{\pi_k}(s,a) + c\big(\log \pi_k(a|s) - \log \pi_{\mathrm{adv}}(a|s)\big) - \alpha \log \pi(a|s)\right] \\
&= \mathbb{E}_s \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_{\pi_k}(s,a) + \boxed{c\big(\log \pi_k(a|s) - \log \pi(a|s)} + \boxed{\log \pi(a|s) - \log \pi_{\mathrm{adv}}(a|s)}\big) - \alpha \log \pi(a|s)\right] \\
&= \mathbb{E}_s\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_{\pi_k}(s,a)\right] \underbrace{\boxed{- cD_{\mathrm{KL}}\big(\pi(\cdot|s)\|\pi_k(\cdot|s)\big)}}_{\pi_k \text{ is attractive}} + \underbrace{\boxed{cD_{\mathrm{KL}}\big(\pi(\cdot|s)\|\pi_{\mathrm{adv}}(\cdot|s)\big)}}_{\pi_{\mathrm{adv}} \text{ is repulsive}} \underbrace{+ \alpha\mathcal{H}\big(\pi(\cdot|s)\big)}_{\text{enforces stochastic policies}}\right]
\end{aligned}
$$

Building Motivation

- Thus, in the PI scheme, AGAC finds a policy that maximizes Q-values, while at the same time remaining close to the current policy and far from a mixture of the previous policies (i.e., $\pi_{k-1}, \pi_{k-2}, \pi_{k-3}, \ldots$).
- Note that we experimentally observe that our method performs better with a smaller learning rate for the adversarial network than that of the other networks, which could imply that a stable repulsive term is beneficial.

Building Motivation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- This optimization problem is strongly concave in π (thanks to the entropy term), and is state-wise a Legendre-Fenchel transform. Its solution is given by:

$$\pi_{k+1} \propto \left(\frac{\pi_k}{\pi_{\text{adv}}}\right)^{\frac{c}{\alpha}} \exp \frac{Q_{\pi_k}}{\alpha}$$

- This result gives us some insight into the behavior of the objective function. Notably, in our example, if π_{adv} is fixed and $c = \alpha$, we recover a KL-regularized PI scheme (Geist et al., 2019) with the modified reward $r - c \log \pi_{\text{adv}}$.

Proof

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}} \mathcal{J}_{\mathrm{PI}}(\pi) \propto \left(\frac{\pi_k}{\pi_{\mathrm{adv}}}\right)^{\frac{c}{\alpha}} \exp \frac{Q_{\pi_k}}{\alpha}$$

with the objective function:

$$\mathcal{J}_{\mathrm{PI}}(\pi) = \mathbb{E}_s \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q_{\pi_k}(s,a) + c\left(\log \pi_k (a|s) - \log \pi_{\mathrm{adv}} (a|s)\right) - \alpha \log \pi(a|s) \right]$$

Proof

- We first consider a simpler optimization problem:

$$\underset{\pi}{\operatorname{argmax}} \langle \pi, Q_{\pi_k} \rangle + \alpha \mathcal{H}(\pi)$$

whose solution is known (Vieillard et al., 2020a, Appendix A).

- The expression for the maximizer is the α-scaled softmax:

$$\pi^* = \frac{\exp\left(\frac{Q_{\pi_k}}{\alpha}\right)}{\left\langle 1, \exp\left(\frac{Q_{\pi_k}}{\alpha}\right)\right\rangle}$$

Proof

- We now turn towards the optimization problem of interest, which we can rewrite as:

$$\operatorname*{argmax}_{\pi} \langle \pi, Q_{\pi_k} + c(\log \pi_k - \log \pi_{\text{adv}}) \rangle + \alpha \mathcal{H}(\pi)$$

- By the simple change of variable $\tilde{Q}_{\pi_k} = Q_{\pi_k} + c(\log \pi_k - \log \pi_{\text{adv}})$, we can reuse the previous solution (replacing Q_{π_k} by $\tilde{Q}_{\pi_k}$).
- With the simplification:

$$\exp \frac{Q_{\pi_k} + c(\log \pi_k - \log \pi_{\text{adv}})}{\alpha} = \left(\frac{\pi_k}{\pi_{\text{adv}}}\right)^{\frac{c}{\alpha}} \exp \frac{Q_{\pi_k}}{\alpha}$$

Implementation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- In all of the experiments, we use PPO (Schulman et al., 2017) as the base algorithm and build on it to incorporate our method. Hence,

In PPO, $L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\,\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$

$$\mathcal{L}_{PG} = -\frac{1}{N}\sum_{t'=t}^{t+N} \min\left(\frac{\pi(a_{t'}|s_{t'},\theta)}{\pi(a_{t'}|s_{t'},\theta_{\text{old}})} A_{t'}^{\text{AGAC}}, \mathrm{clip}\left(\frac{\pi(a_{t'}|s_{t'},\theta)}{\pi(a_{t'}|s_{t'},\theta_{\text{old}})}, 1-\epsilon, 1+\epsilon\right) A_{t'}^{\text{AGAC}}\right)$$

with $A_{t'}^{\text{AGAC}}$ given in first equation, N the temporal length considered for one update of parameters and ϵ the clipping parameter.

Implementation

- Similar to RIDE (Raileanu & Rocktäschel, 2019), we also discount PPO by episodic state visitation counts.
- The actor, critic and adversary use the convolutional architecture of the Nature paper of DQN (Mnih et al., 2015) with different hidden sizes.
- The three neural networks are optimized using Adam (Kingma & Ba, 2015). Our method does not use RNNs in its architecture; instead, in all our experiments, we use frame stacking.

Implementation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

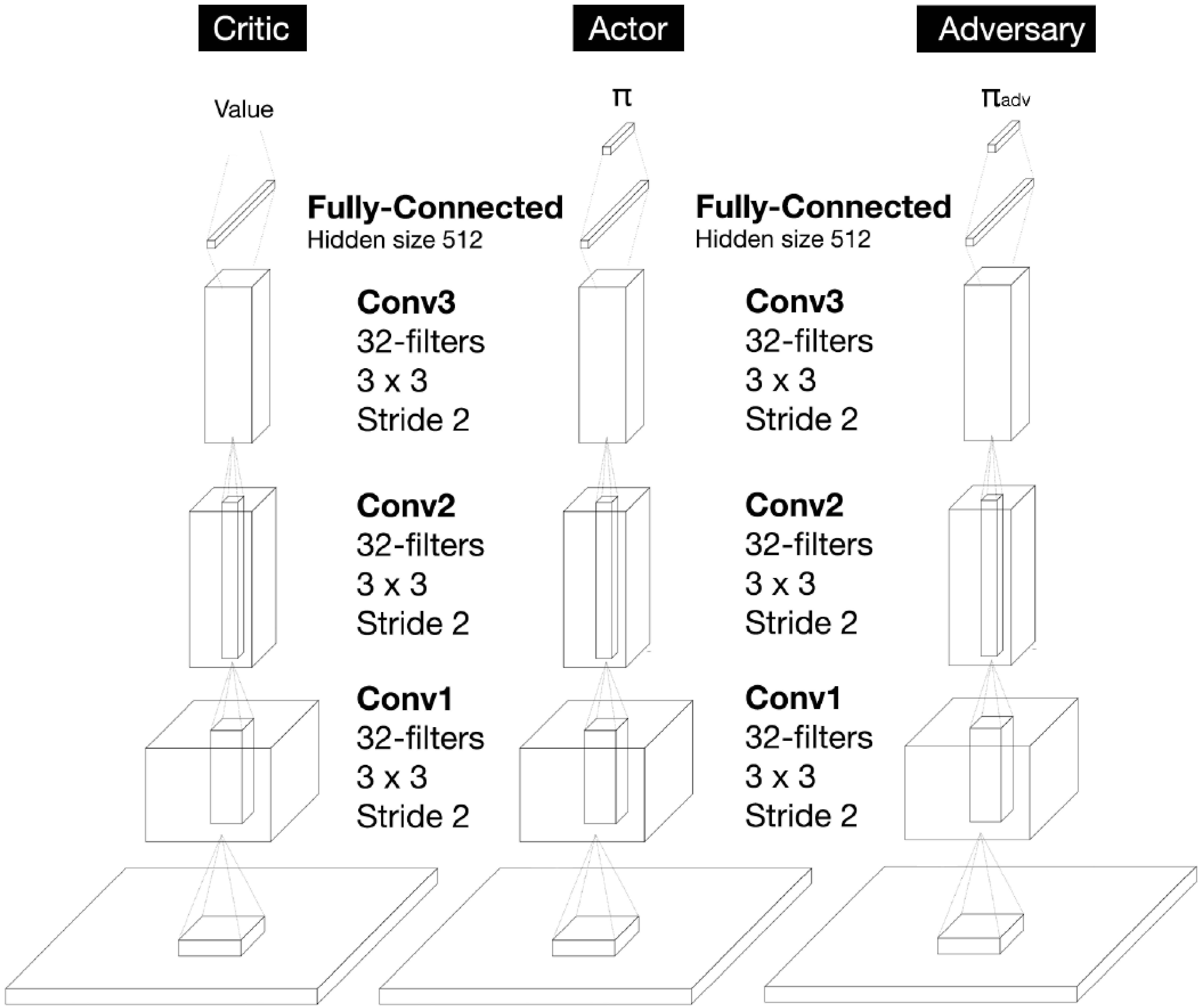


Figure 11: Artificial neural architecture of the critic, the actor and the adversary.

Implementation

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- At each training step, we perform a stochastic optimization step to minimize $\mathcal{L}_{\text{AGAC}}$ using stop-gradient:
  - $\theta \leftarrow \text{Adam}(\theta, \nabla_\theta \mathcal{L}_{\text{PG}}, \eta_1)$
  - $\phi \leftarrow \text{Adam}(\phi, \nabla_\phi \mathcal{L}_{\text{V}}, \eta_1)$
  - $\psi \leftarrow \text{Adam}(\psi, \nabla_\psi \mathcal{L}_{\text{adv}}, \eta_2)$

Table 3: Hyperparameters used in AGAC.

Parameter	Value
Horizon T	2048
Nb. epochs	4
Nb. minibatches	8
Nb. frames stacked	4
Nonlinearity	ELU (Clevert et al., 2016)
Discount γ	0.99
GAE parameter λ	0.95
PPO clipping parameter ϵ	0.2
β_V	0.5
c	$4 \cdot 10^{-4}$ ($4 \cdot 10^{-5}$ in VizDoom)
c anneal schedule	linear
β_{adv}	$4 \cdot 10^{-5}$
Adam stepsize η_1	$3 \cdot 10^{-4}$
Adam stepsize η_2	$9 \cdot 10^{-5} = 0.3 \cdot \eta_1$

Experiments

- In this section, we describe our experimental study in which we investigate:
  - (i) whether the adversarial bonus alone (e.g. without episodic state visitation count) is sufficient to outperform other methods in VizDoom, a sparse-reward task with high-dimensional observations
  - (ii) whether AGAC succeeds in partially-observable and procedurally-generated environments with high sparsity in the rewards, compared to other methods

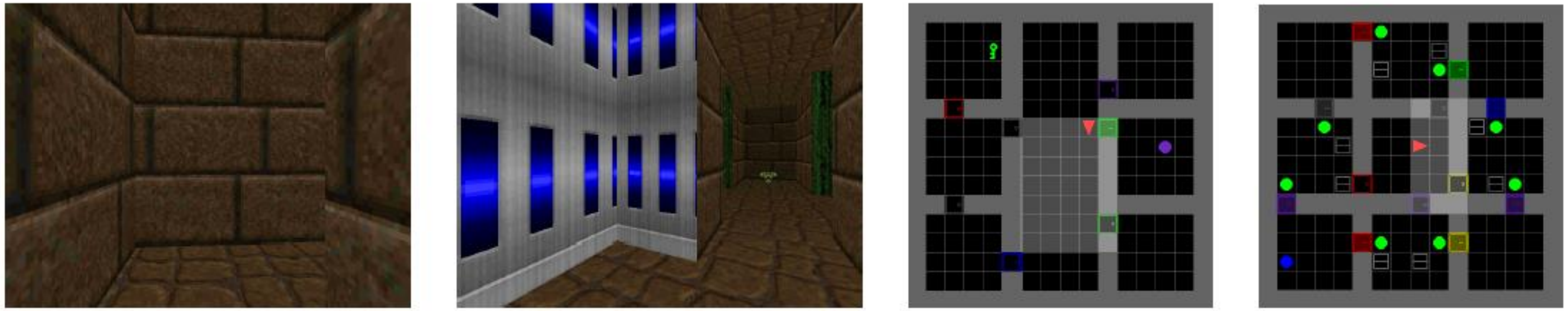
Experiments

- In this section, we describe our experimental study in which we investigate:
  - (iii) how well AGAC is capable of exploring in environments without extrinsic reward
  - (iv) the training stability of our method. In all of the experiments, lines are average performances and shaded areas represent one standard deviation

Experiments

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Environments



(a) (b) (c) (d)

Figure 1: (a,b) Frames from the 3-D navigation task VizdoomMyWayHome. (c) MiniGrid-KeyCorridorS6R3. (d) MiniGrid-ObstructedMazeFull.

Experiments

- Baselines
  - RIDE (Raileanu & Rocktäschel, 2019)
  - Count as Count-Based Exploration (Bellemare et al., 2016b)
  - RND (Burda et al., 2018)
  - ICM (Pathak et al., 2017)
  - AMIGo (Campero et al., 2021)

Experiments

- Adversarially-based Exploration (No episodic count)

Table 1: Average return in VizDoom at different timesteps.

Nb. of Timesteps	2M	4M	6M	8M	10M
AGAC	$\mathbf{0.74} \pm 0.05$	$\mathbf{0.96} \pm 0.001$	$\mathbf{0.96} \pm 0.001$	$\mathbf{0.97} \pm 0.001$	$\mathbf{0.97} \pm 0.001$
RIDE	0.	0.	0.95 ± 0.001	$\mathbf{0.97} \pm 0.001$	$\mathbf{0.97} \pm 0.001$
ICM	0.	0.	0.95 ± 0.001	$\mathbf{0.97} \pm 0.001$	$\mathbf{0.97} \pm 0.001$
AMIGo	0.	0.	0.	0.	0.
RND	0.	0.	0.	0.	0.
Count	0.	0.	0.	0.	0.

Experiments

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Hard-Exploration Tasks with Partially-Observable Environments

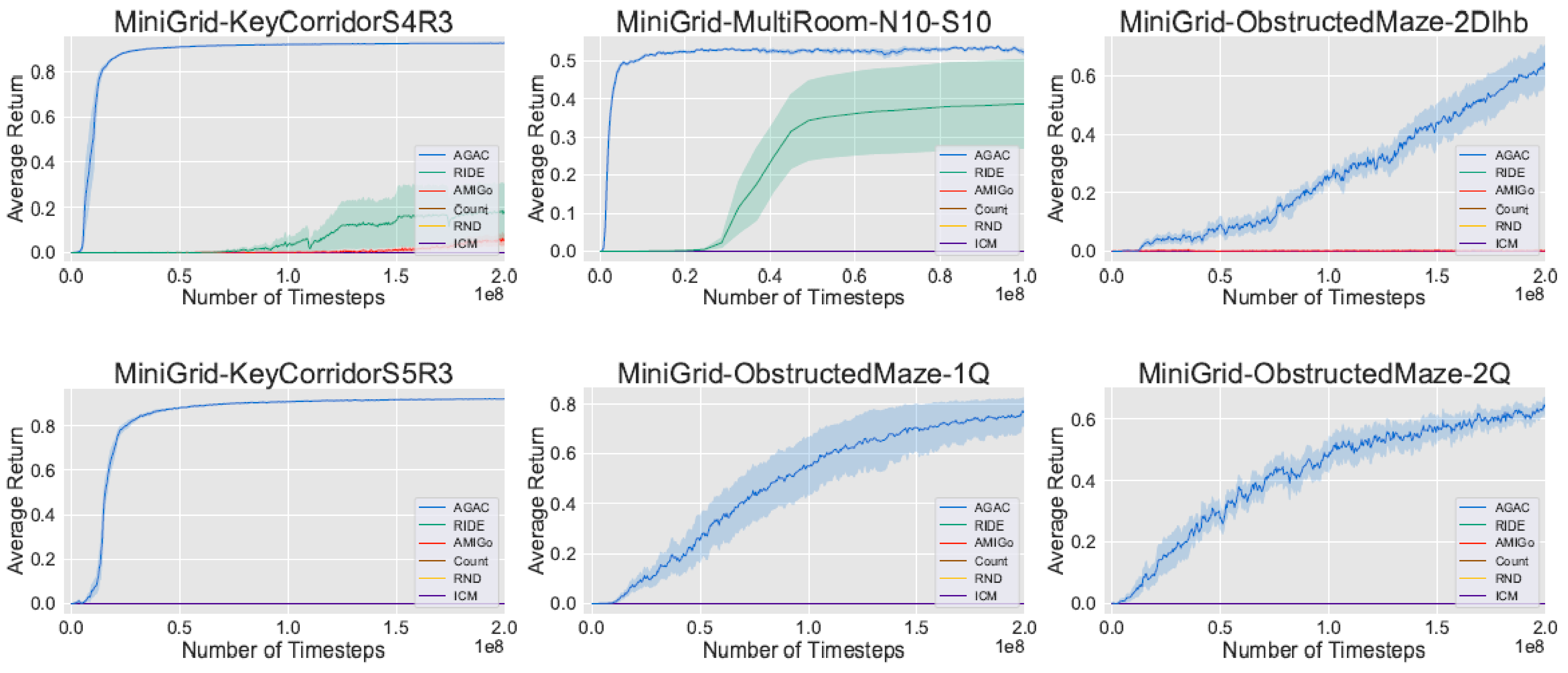


Figure 2: Performance evaluation of AGAC.

Experiments

- Training Stability

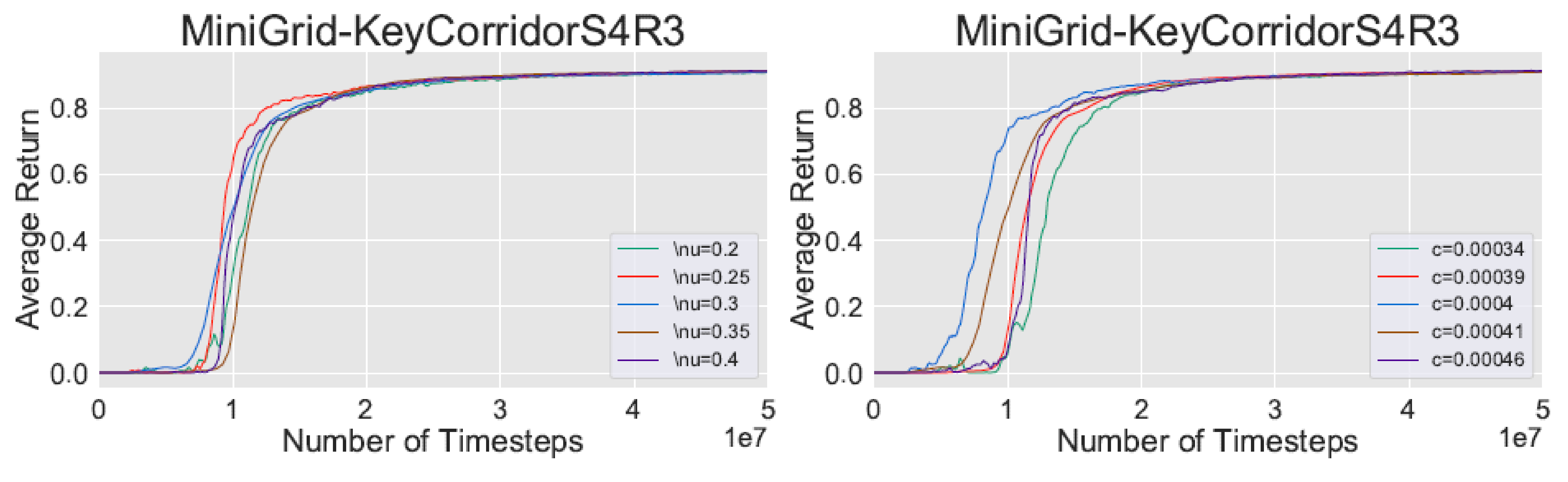


Figure 3: Sensitivity analysis of AGAC in KeyCorridorS4R3.

Experiments

Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Exploration in Reward-Free Environment

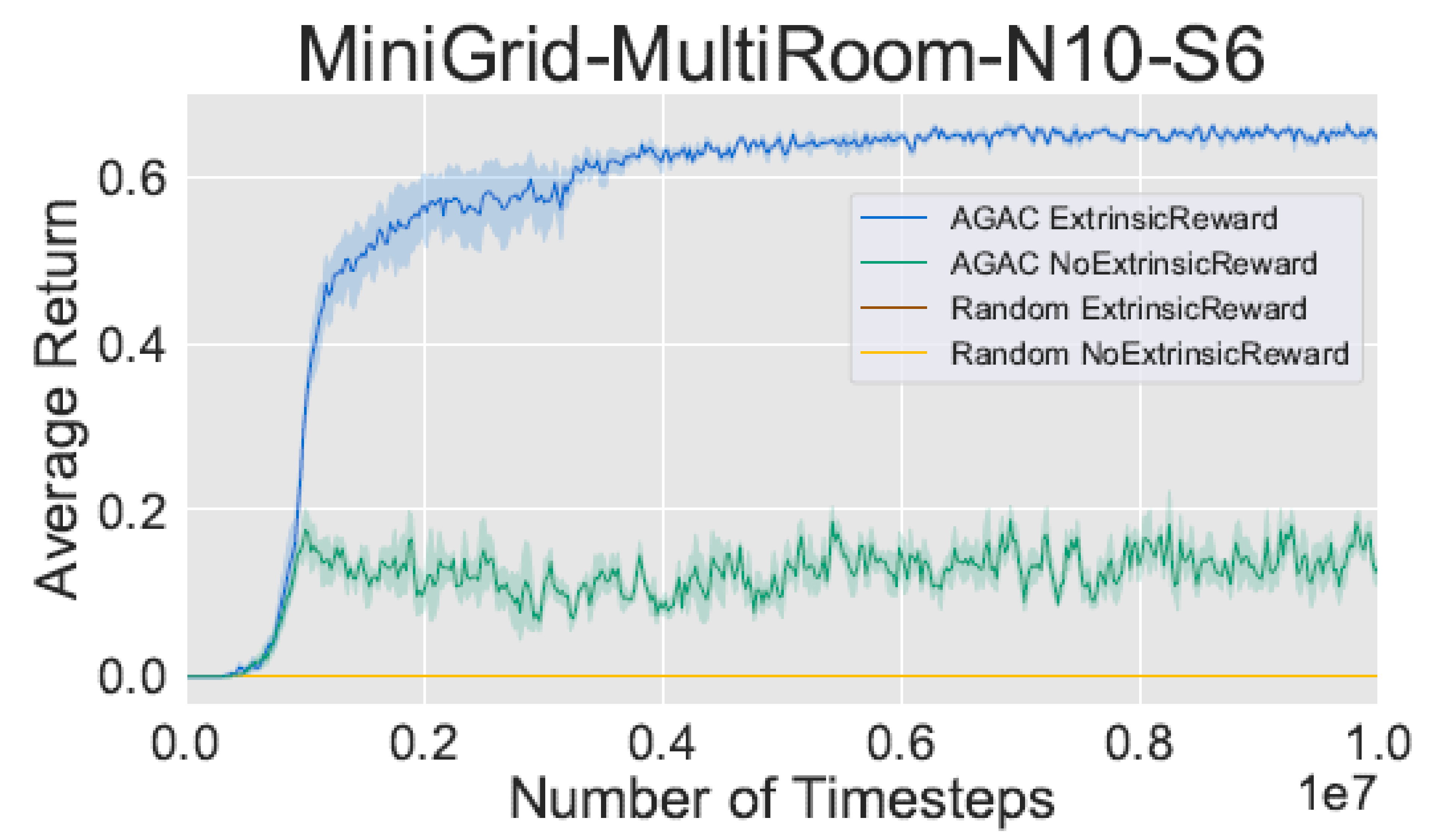
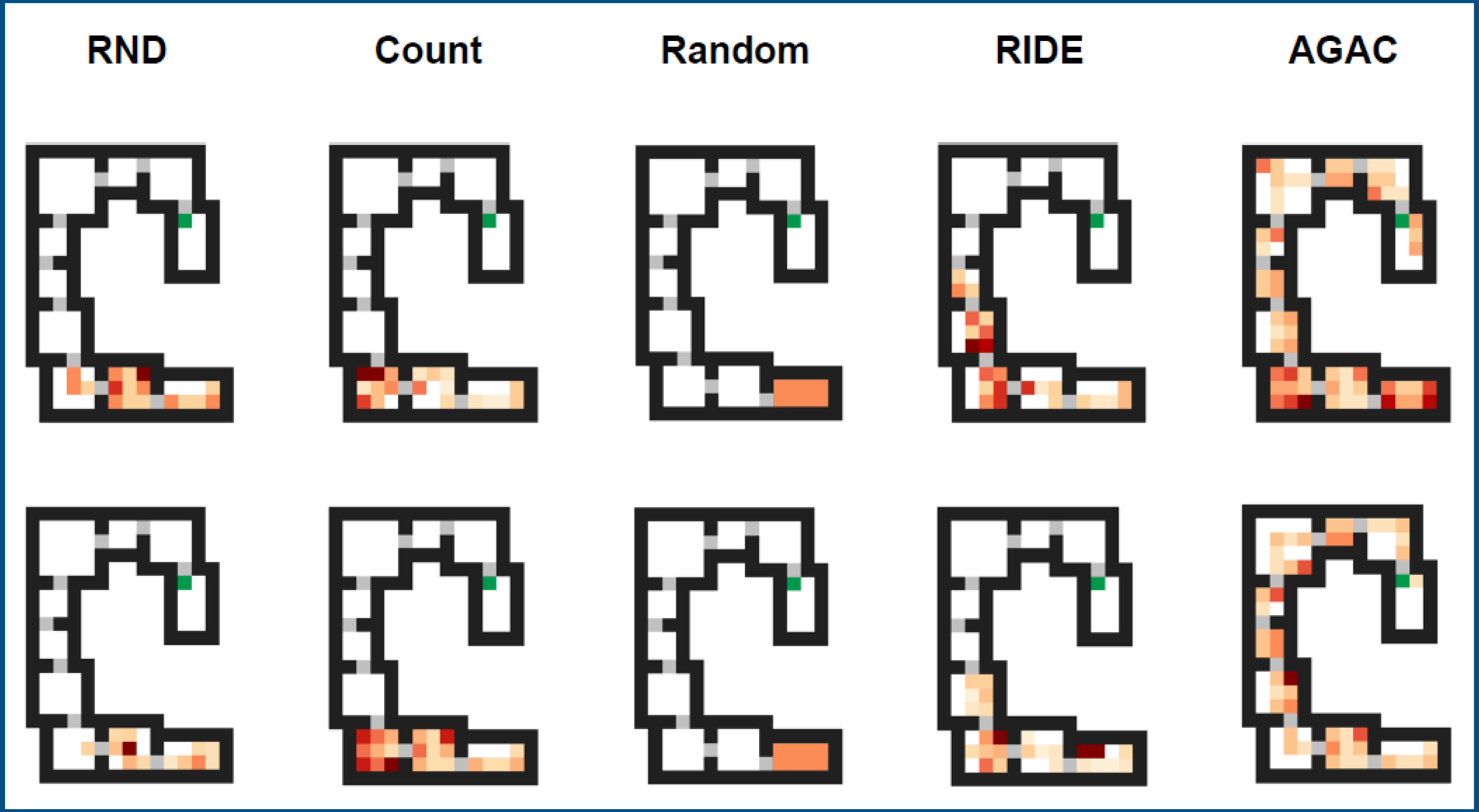


Figure 4: Average return on N10S6 with and without extrinsic reward.

Experiments

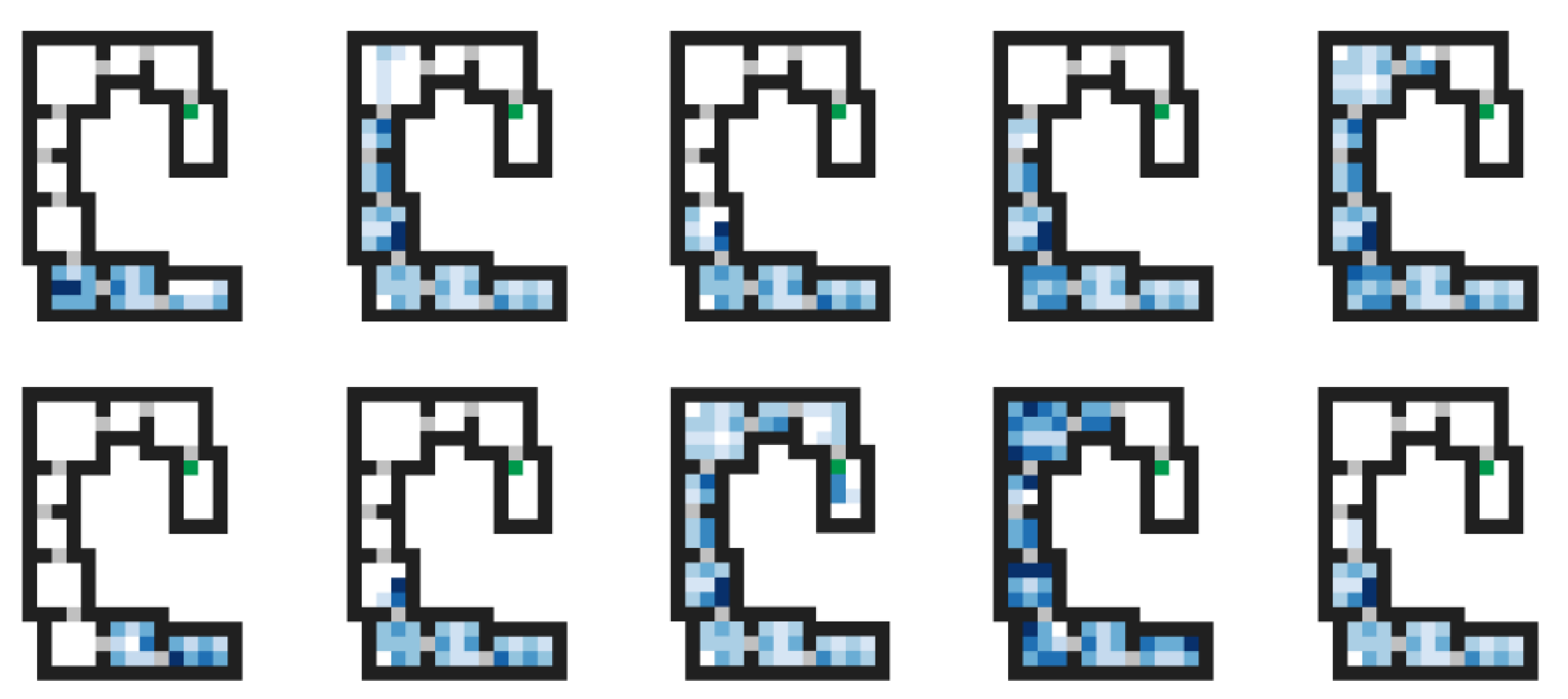
Adversarially Guided Actor-Critic,
Y. Flet-Berliac et al, 2021

- Visualizing Coverage and Diversity



Experiments

- Visualizing Coverage and Diversity



Disscussions

- This paper introduced AGAC, a modification to the traditional actor-critic framework: an adversary network is added as a third protagonist.
- The mechanics of AGAC have been discussed from a policy iteration point of view, and we provided theoretical insight into the inner workings of the proposed algorithm: the adversary forces the agent to remain close to the current policy while moving away from the previous ones. In a nutshell, the influence of the adversary makes the actor <u>conservatively diversified</u>.

Thank you!