Adversarially Guided Actor-Critic

Yannis Flet-Berliac*[1,2], based on work with:
Johan Ferret*[1,2,3], Olivier Pietquin[3], Philippe Preux[1,2], Matthieu Geist[3]

[1]Inria, SequeL team
[2]Univ. Lille, CRIStAL, CNRS
[3]Google Research, Brain team
*Equal contribution

DeepMind Reading Group

February 12, 2021

Agenda

- Policy Gradients (PG) and Actor-Critic (AC) methods
  - Contextualization
  - Critics in deep PG algorithms
- Problem: popular AC methods *fail* ...
  - ... where efficient exploration is a bottleneck
  - ... to generalize correctly [Song et al., 2020, Cobbe et al., 2020]
- `AGAC`: an adversary to make the agent *conservatively diversified*
  - Adding an adversary network as a third component to the AC framework
  - Building motivation from a PI point of view
- `AGAC`: how well does it work?
  - Adversarially-based exploration: VizDoom
  - Hard-exploration tasks with partially-observable environments
  - Investigating trajectory coverage and strategy diversity
- Conclusion and perspectives

Reinforcement Learning

Environment (Markov Decision Process):

- State $s \in \mathcal{S}$, action $a \in \mathcal{A}$
- Reward function: $r(s, a)$, transition probabilities: $P(s'|s, a)$

Agent:

- Stochastic policy $\pi_\theta(a|s)$ with parameter θ

An agent in state s_t interacts with an environment by sampling action $a_t \sim \pi_\theta(\cdot|s_t)$, receives reward r_t and transitions to a new state s_{t+1}.

Reinforcement Learning

Environment (Markov Decision Process):

- State $s \in \mathcal{S}$, action $a \in \mathcal{A}$
- Reward function: $r(s, a)$, transition probabilities: $P(s'|s, a)$

Agent:

- Stochastic policy $\pi_\theta(a|s)$ with parameter θ

An agent in state s_t interacts with an environment by sampling action $a_t \sim \pi_\theta(\cdot|s_t)$, receives reward r_t and transitions to a new state s_{t+1}.

Goal: Find π that maximizes

$$J(\pi_\theta) \triangleq \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right)\right]$$

with $\gamma \in [0, 1)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, $a_t \sim \pi_\theta(\cdot|s_t)$ and trajectory τ.

Policy Gradients

Policy gradient algorithms try to solve the optimization problem

$$\max_{\theta} J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right)\right]$$

by taking stochastic gradient ascent on the policy parameters θ, using the policy gradient

$$\nabla_\theta J = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t, a_t)\right]$$

with $Q^{\pi_\theta}(s, a) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right)|s_0 = s, a_0 = a\right]$.

Intuition: make the good actions more probable.

Policy Gradients

It is possible to obtain an unbiased estimate of the policy gradient from empirical trajectories ...

... But the corresponding **variance can be extremely high**.

Policy Gradients

It is possible to obtain an unbiased estimate of the policy gradient from empirical trajectories ...

... But the corresponding **variance can be extremely high**.

Subtracting a baseline from the value function in the policy gradient can be very beneficial in reducing variance without damaging the bias [Williams, 1992, Weaver and Tao, 2001].

Policy Gradients

It is possible to obtain an unbiased estimate of the policy gradient from empirical trajectories ...

... But the corresponding **variance can be extremely high**.

Subtracting a baseline from the value function in the policy gradient can be very beneficial in reducing variance without damaging the bias [Williams, 1992, Weaver and Tao, 2001].

In practice, if we denote $\hat{X}$ the empirical estimate of X, the policy gradient becomes

$$\nabla_\theta J = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t | s_t) \hat{A}^{\pi_\theta}(s_t, a_t) \right],$$

with $\hat{A}^{\pi_\theta}(s, a) = \hat{Q}^{\pi_\theta}(s, a) - \hat{V}^{\pi_\theta}(s)$ the advantage estimate which quantifies how an action a is better than the average action in state s.

Critics in Deep Policy Gradients

$\hat{V}^{\pi_\theta}$ is learned using a function estimator.

Let $V_\phi : \mathcal{S} \to \mathbb{R}$ (ϕ its parameter) be an estimator of the empirical return $\hat{V}^{\pi_\theta}$. V_ϕ is traditionally learned through minimizing the MSE against $\hat{V}^{\pi_\theta}$. The critic minimizes:

$$\mathcal{L}_{\mathsf{V}} = \mathbb{E}_s \left[\left(V_\phi(s) - \hat{V}^{\pi_{\theta_{\text{old}}}}(s) \right)^2 \right],$$

where the states s are collected under policy $\pi_{\theta_{\text{old}}}$ at the previous iteration.

$\to$ V_ϕ is called **the critic**.

Setting applicable to e.g. PPO [Schulman et al., 2017].

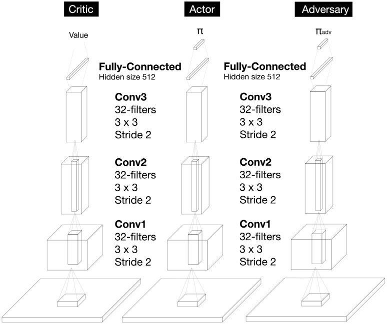
AGAC: a new protagonist to the actor-critic setting

In AGAC, the **adversary policy** π_{adv} mimics the actor policy π:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_s \left[D_{\mathrm{KL}}(\pi(\cdot|s, \theta_{\text{old}}) \| \pi_{\text{adv}}(\cdot|s, \psi)) \right]$$

with ψ the parameters of π_{adv} and θ_{old} that of π at the previous iteration.

→ The **adversary** tries to predict the actions of the actor.



AGAC: a new protagonist to the actor-critic setting

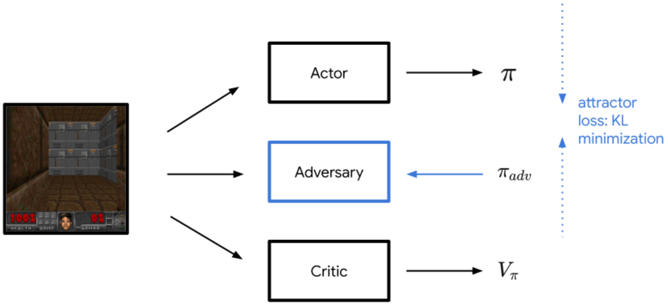


Figure: The adversary minimizes the discrepancy between its action distribution π_{adv} and the distribution induced by the policy π.

AGAC: a new protagonist to the actor-critic setting

AGAC modifies the AC **advantage and value functions**:

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

$$\mathcal{L}_V = \mathbb{E}_s \left[\left(V_\phi(s) - \left(\hat{V}^{\pi_{\theta_{\text{old}}}}(s) + c\, D_{\mathrm{KL}} \left(\pi(\cdot|s, \theta_{\text{old}}) \| \pi_{\text{adv}}(\cdot|s, \psi_{\text{old}}) \right) \right) \right)^2 \right]$$

with c a varying hyperparameter.

→ The **actor** (a) maximizes the sum of expected returns;
(b) counteracts the adversary's predictions.

AGAC: a new protagonist to the actor-critic setting

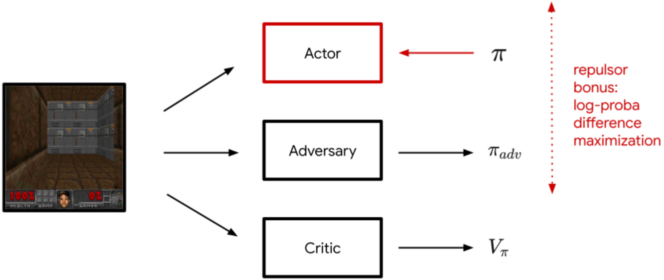


Figure: The actor counteracts the adversary's predictions by maximizing the discrepancy between π and π_{adv} (in addition to finding the optimal actions to maximize the sum of expected returns).

AGAC: final objective function

AGAC minimizes the following loss:

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

Building motivation

In the PI scheme, AGAC would modify the action-value as:

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

with π_k the policy at iteration k.

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

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

Idea: we can rewrite this objective

$$\mathcal{J}_{\text{PI}}(\pi) = \mathbb{E}_s \Big[\mathbb{E}_{a\sim\pi(\cdot|s)}[Q_{\pi_k}(s,a)] \underbrace{- c\, D_{\mathrm{KL}}(\pi(\cdot|s)||\pi_k(\cdot|s))}_{\pi_k \text{ is attractive}} \underbrace{+ c\, D_{\mathrm{KL}}(\pi(\cdot|s)||\pi_{\text{adv}}(\cdot|s))}_{\pi_{\text{adv}} \text{ is repulsive}} \underbrace{+ \alpha\, \mathcal{H}(\pi(\cdot|s))}_{\text{enforces stochastic policies}} \Big].$$

Building motivation

$$\mathcal{J}_{\text{PI}}(\pi) = \mathbb{E}_s\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}[Q_{\pi_k}(s,a)] \underbrace{- c\, D_{\mathrm{KL}}(\pi(\cdot|s)||\pi_k(\cdot|s))}_{\pi_k \text{ is attractive}} \underbrace{+ c\, D_{\mathrm{KL}}(\pi(\cdot|s)||\pi_{\text{adv}}(\cdot|s))}_{\pi_{\text{adv}} \text{ is repulsive}} \underbrace{+ \alpha\, \mathcal{H}(\pi(\cdot|s))}_{\text{enforces stochastic policies}}\Big].$$

$\rightarrow$ AGAC finds a policy that:

(a) maximizes Q-values;
(b) remains close to the current policy;
(c) remains far from a mixture of previous policies (*i.e.*, $\pi_{k-1}, \pi_{k-2}, \ldots$).

The actor's policy is *conservatively diversified*.

Empirical results: adversarially-based exploration

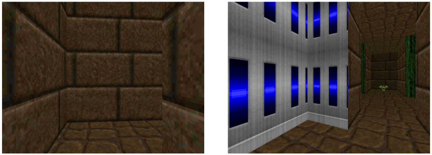


Figure: Frames from the 3-D navigation task *VizdoomMyWayHome*.

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.

Importantly, other algorithms benefit from **count-based exploration**.

Empirical results: procedurally-generated partially-observable environments

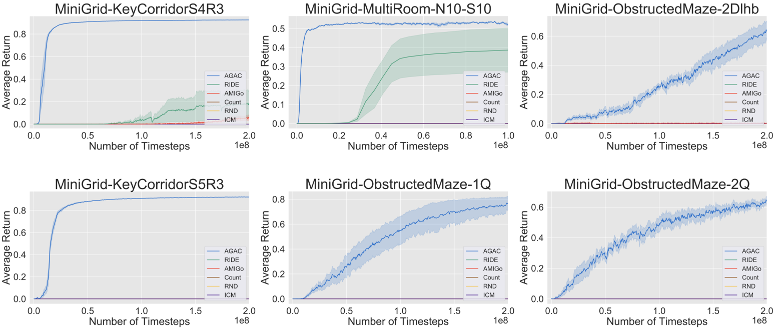


Figure: Performance evaluation of AGAC on MiniGrid tasks.

Here, `AGAC` also uses **episodic state visitation counts**.

Empirical results: some insights

Two main arguments to explain why `AGAC` is successful:

- the exploration bonus does not **dissipate** compared to most other methods (see Fig. 9 [Flet-Berliac et al., 2021]);
- `AGAC` does not make **assumptions about the environment dynamics** (*e.g.* RIDE [Raileanu and Rocktäschel, 2019] assume changes in the environment following an action).

Empirical results: visualizing exploration coverage

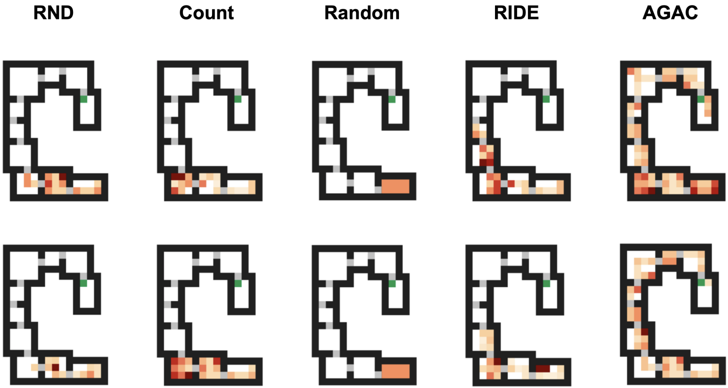


Figure: State visitation heatmaps for different methods trained in a singleton environment (top row) and procedurally-generated environments (bottom row) without extrinsic reward for 10M timesteps in the *MultiRoomN10S6* task.

Conclusion and perspectives

This paper ...:

◦ introduces a modification of the traditional actor-critic framework which (a) is motivated from a theoretical standpoint using the (simplified) point of view of PI (b) produces considerable gains in performance

◦ highlights the benefits of a more extended investigtion of count-less methods for hard-exploration and procedurally-generated tasks

◦ is not a claim that `AGAC` is the best version of the proposed "adversarially guided AC" formulation *i.e.* many components could be improved (better NN architecture, Polyak averaging, etc.)

◦ could be followed-up with further analysis of the adversarial bonus (although our training stability study indicates that c is $+$ sensitive than other HP, why not try with a dynamic c)

◦ could be extended to stochastic environments

Thank you!

Questions?

Flet-Berliac, Y., Ferret, J., Pietquin, O., Preux, P., and Geist, M. (2021). **Adversarially guided actor-critic**. In *International Conference on Learning Representations*.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. (2020). **Leveraging procedural generation to benchmark reinforcement learning**. In *International Conference on Machine Learning*.

Puterman, M. (1994). **Markov Decision Processes**. *John Wiley & Sons*.

Raileanu, R. and Rocktaschel, T. (2019). **Ride: Rewarding impact-driven exploration for procedurally-generated environments**. In *International Conference on Learning Representations*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). **Proximal policy optimization algorithms**. *arXiv preprint arXiv:1707.06347*.

Song, X., Jiang, Y., Du, Y., and Neyshabur, B. (2020). **Observational overfitting in reinforcement learning**. In *International Conference on Learning Representations*.

Weaver, L. and Tao, N. (2001). **The optimal reward baseline for gradient-based reinforcement learning**. In *Advances in Neural Information Processing Systems*.

More results: visualizing exploration coverage

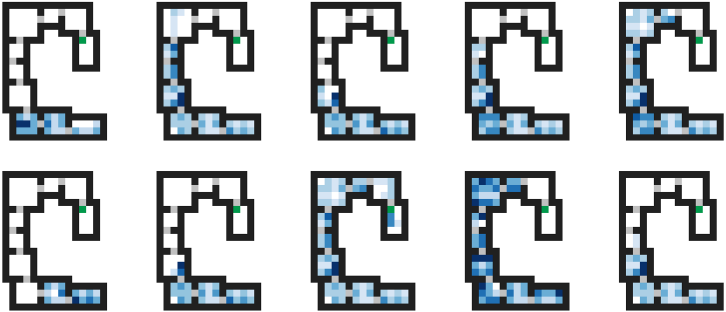


Figure: State visitation heatmaps of the last ten episodes of an agent trained in procedurally- generated environments without extrinsic reward for 10M timesteps in the *MultiRoomN10S6* task. The agent is continuously engaging in new strategies.

More results: visualizing exploration coverage

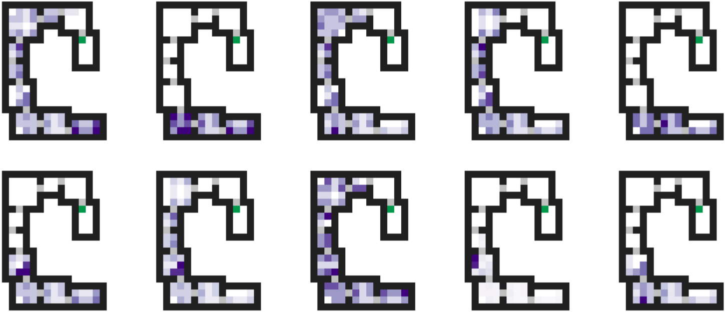


Figure: State visitation heatmaps of the last ten episodes of an agent trained in a singleton environment with no extrinsic reward 10M timesteps in the *MultiRoomN10S6* task. The agent is continuously engaging into new strategies.

More results: reward free

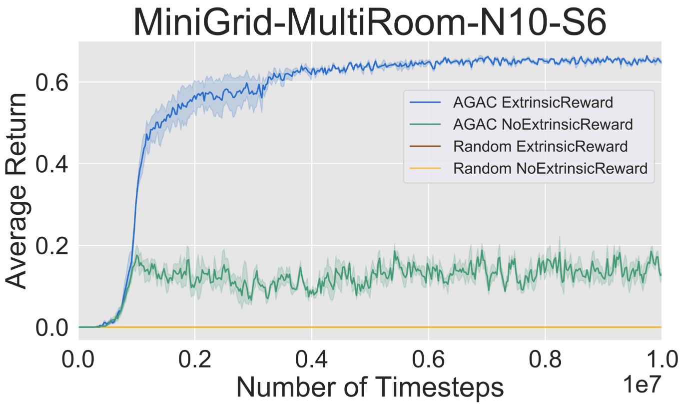


Figure: Average return on N10S6 with and without extrinsic reward.

More results: intrinsic reward

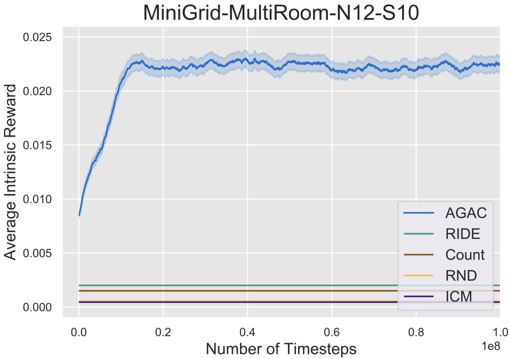


Figure: Average intrinsic reward for different methods trained in *MultiRoomN12S10*.

More results: training stability

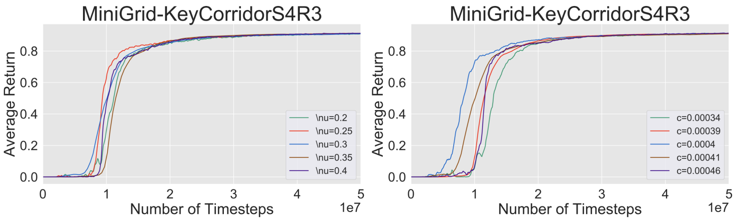


Figure: Sensitivity analysis of AGAC in *KeyCorridorS4R3*.

More results: extremely hard-exploration tasks

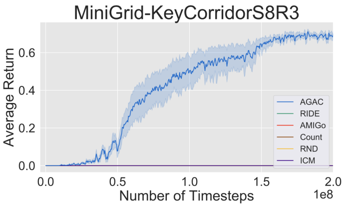


Figure: Performance evaluation of AGAC compared to RIDE, AMIGo, Count, RND and ICM on extremely hard-exploration problems.