~~Generalized~~ Off-Policy Actor-Critic

University of Oxford
Shangtong Zhang, Wendelin Boehmer, Shimon Whiteson

Off-Policy Actor-Critic optimizes the excursion objective

The excursion objective (Degris et al, 2012): $$J_\mu \doteq \sum_s d_\mu(s) v_\pi(s)$$

- Off-PAC (Degris et al, 2012)
- (off-policy) DPG (Silver et al, 2014)
- DDPG (Lillicrap et al, 2015)
- ACER (Wang et al, 2016)
- (off-policy) EPG (Ciosek and Whiteson, 2017)
- IMPALA (Espeholt et al, 2018)
- EAC (Maei, 2018)
- ACE (Imani et al, 2018)

μ : **behavior policy**

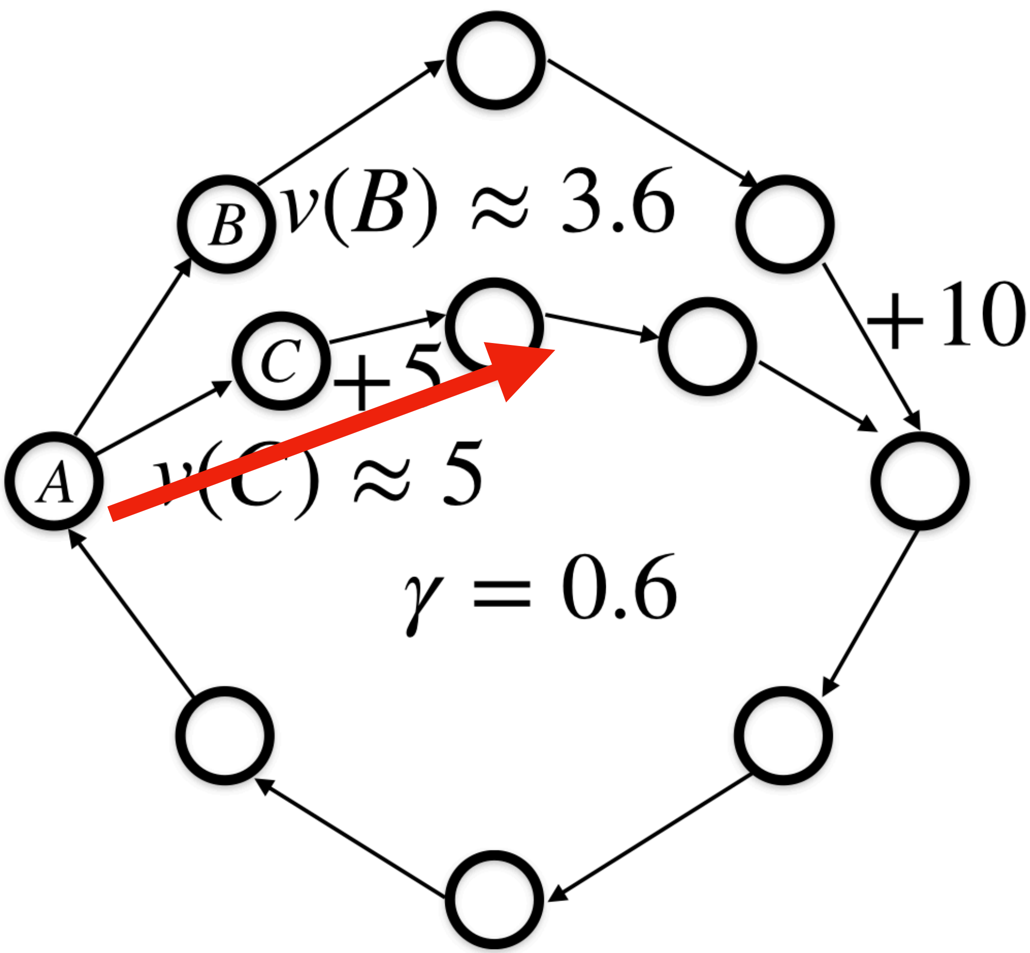
π : **target policy**

d_μ : **steady distribution**

?

The excursion objective leads to an inferior solution

A two-circle MDP:

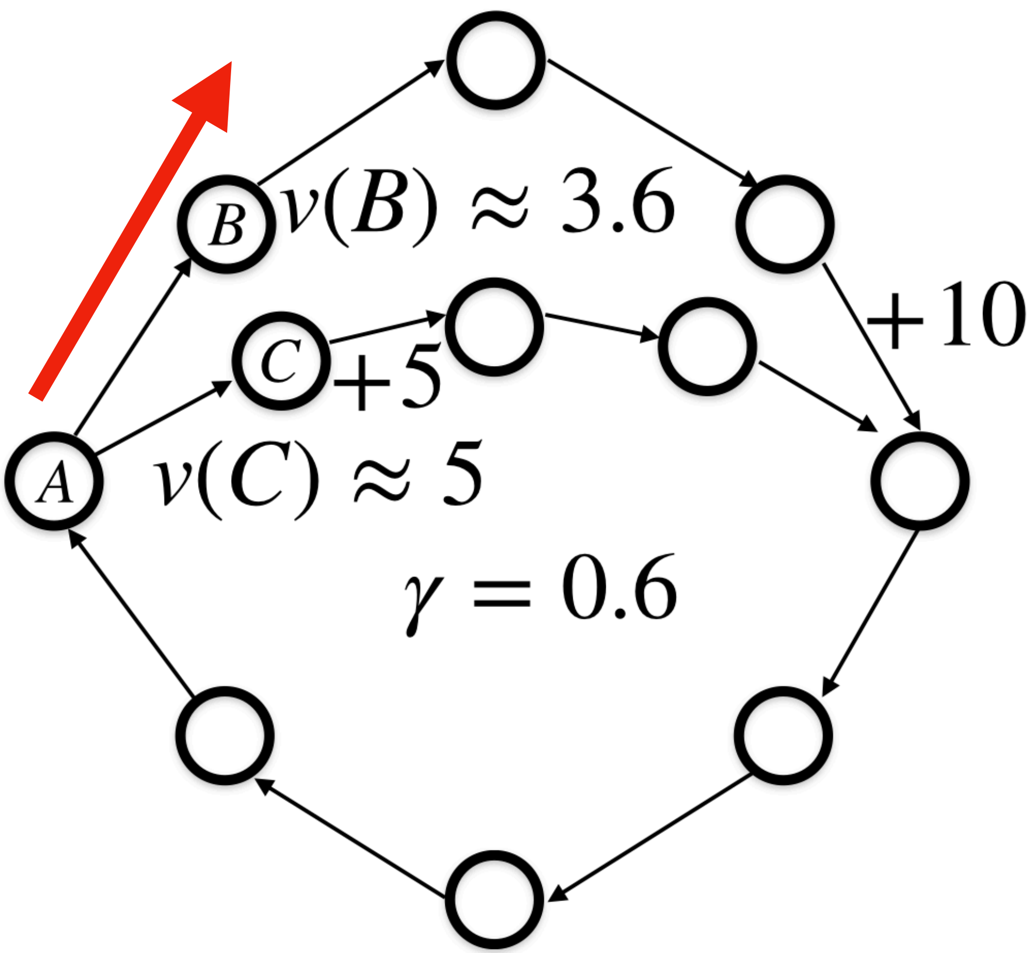


$$J_\mu \doteq \sum_s d_\mu(s) v_\pi(s)$$

μ : **random**

The alternative life objective leads to the optimal solution

A two-circle MDP:



The alternative life objective:

$$J_\pi \doteq \sum_s d_\pi(s) v_\pi(s)$$

(Average reward)

We want to unify the alt-life objective and the excursion objective

$$J_\mu \doteq \sum_s d_\mu(s) v_\pi(s)$$

biased but easy to optimize

$$J_\pi \doteq \sum_s d_\pi(s) v_\pi(s)$$

unbiased but hard to optimize

$TD(0)$ $\quad$ $TD(\lambda)$ $\quad$ $Monte\ Carlo$

Let's start with a unifying transition matrix

$$\mathbf{P}_{\hat{\gamma}} \doteq \hat{\gamma}\mathbf{P}_{\pi} + (1-\hat{\gamma})\mathbf{1}\mathbf{d}_{\mu}^{\top} \quad (\hat{\gamma} \in [0,1])$$

$$\mathbf{d}_{\hat{\gamma}} = (1-\hat{\gamma})(\mathbf{I} - \hat{\gamma}\mathbf{P}_{\pi}^{\top})^{-1}\mathbf{d}_{\mu} \quad (\hat{\gamma} < 1)$$

$$\mathbf{d}_{\hat{\gamma}} = \mathbf{d}_{\pi} \quad (\hat{\gamma} = 1)$$

$$\lim_{\hat{\gamma}\to 1} \mathbf{d}_{\hat{\gamma}} \overset{?}{\neq} \mathbf{d}_{\pi}$$

Gelada and Bellemare, 2019

The counterfactual objective unifies two old objectives

$$J_\pi \doteq \sum_s d_\pi(s) v_\pi(s) \qquad\qquad J_\mu \doteq \sum_s d_\mu(s) v_\pi(s)$$

The counterfactual objective:

$$J_{\hat{\gamma}} \doteq \sum_s d_{\hat{\gamma}}(s) v_\pi(s)$$

$$\lim_{\hat{\gamma} \to 1} J_{\hat{\gamma}} \stackrel{?}{\doteq} J_\pi$$

The counterfactual objective converges to the alt-life objective

$$\Pi_{\hat{\gamma}} \doteq \mathbf{1}\mathbf{d}_{\hat{\gamma}}^{\top} \quad (\hat{\gamma} \in [0,1])$$

$$\lim_{t\to\infty} \mathbf{P}_{\hat{\gamma}}^{t} = \Pi_{\hat{\gamma}}$$ **(Levin et al., 2017)**

$$\lim_{\hat{\gamma}\to 1} \mathbf{1}\mathbf{d}_{\hat{\gamma}}^{\top} = \lim_{\hat{\gamma}\to 1} \lim_{t\to\infty} \mathbf{P}_{\hat{\gamma}}^{t}$$

$$\lim_{t\to\infty} \lim_{\hat{\gamma}\to 1} \mathbf{P}_{\hat{\gamma}}^{t} = \lim_{t\to\infty} \mathbf{P}_{\pi}^{t} = \mathbf{1}\mathbf{d}_{\pi}^{\top}$$

The counterfactual objective converges to the alt-life objective

$$\lim_{t \to \infty} \mathbf{P}_{\hat{\gamma}}^{t} = \Pi_{\hat{\gamma}}, \textit{ uniformly, for } \hat{\gamma} \in (\hat{\gamma}_0, 1], \textit{ where } \hat{\gamma}_0 \in (0,1)$$

$$\Big\downarrow \ \textbf{Moore-Osgood Theorem}$$

$$\lim_{\hat{\gamma} \to 1} \lim_{t \to \infty} \mathbf{P}_{\hat{\gamma}}^{t} = \lim_{t \to \infty} \lim_{\hat{\gamma} \to 1} \mathbf{P}_{\hat{\gamma}}^{t}$$

Let's compute the policy gradient of the counterfactual objective

$$J_{\hat{\gamma}} = \sum_s d_{\hat{\gamma}}(s) v^{\pi}(s) = \sum_s d_{\mu}(s) \frac{d_{\hat{\gamma}}(s)}{d_{\mu}(s)} v_{\pi}(s) = \sum_s d_{\mu}(s) c(s) v_{\pi}(s)$$

$$\nabla J_{\hat{\gamma}} = \sum_s d_{\mu}(s) c(s) \nabla v_{\pi}(s) + \sum_s d_{\mu}(s) \nabla c(s) v_{\pi}(s)$$

set the interest function to c in Imani et al. (2018)

Let's compute the policy gradient of the counterfactual objective

$$\mathbf{c} = \hat{\gamma}\mathbf{D}_\mu^{-1}\mathbf{P}_\pi^\top\mathbf{D}_\mu\mathbf{c} + (1-\hat{\gamma})\mathbf{1}$$ **(Gelada and Bellemare, 2019)**

$$\nabla\mathbf{c} = \hat{\gamma}\mathbf{D}_\mu^{-1}(\mathbf{I} - \hat{\gamma}\mathbf{P}_\pi^\top)^{-1}\nabla\mathbf{P}_\pi^\top\mathbf{D}_\mu\mathbf{c}$$

$$D_\mu \doteq \textbf{diagnal}(d_\mu)$$

Let's sample the policy gradient of the counterfactual objective

$$\nabla \mathbf{c} = \hat{\gamma}\mathbf{D}_\mu^{-1}\boxed{(\mathbf{I} - \hat{\gamma}\mathbf{P}_\pi^\top)^{-1}}\boxed{\nabla \mathbf{P}_\pi^\top \mathbf{D}_\mu \mathbf{c}}$$

ETD (Sutton et al, 2016)

An intrinsic interest function

COP-TD (Hallak and Mannor, 2017)

We arrive at the Generalized Off-Policy Actor-Critic (Geoff-PAC)

$$\nabla J_{\hat{\gamma}} = \sum_s d_\mu(s) c(s) \nabla v_\pi(s) + \sum_s d_\mu(s) \nabla c(s) v_\pi(s)$$

$$F_t^{(1)} \doteq c(S_t) + \gamma \rho_{t-1} F_{t-1}^{(1)}$$

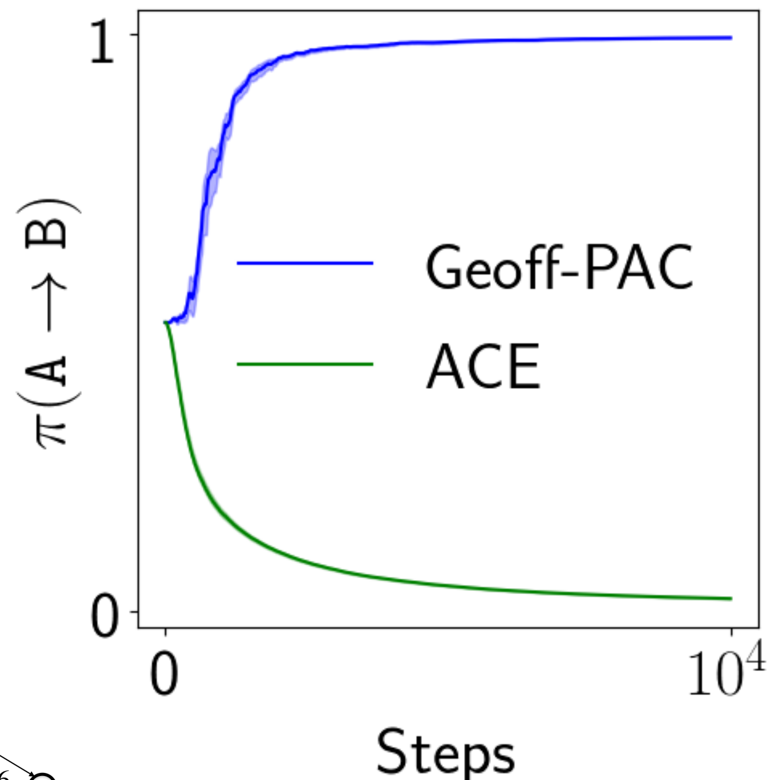
$$Z_t^{(1)} \doteq \rho_t F_t^{(1)} q_\pi(S_t, A_t) \nabla \log \pi(S_t, A_t)$$

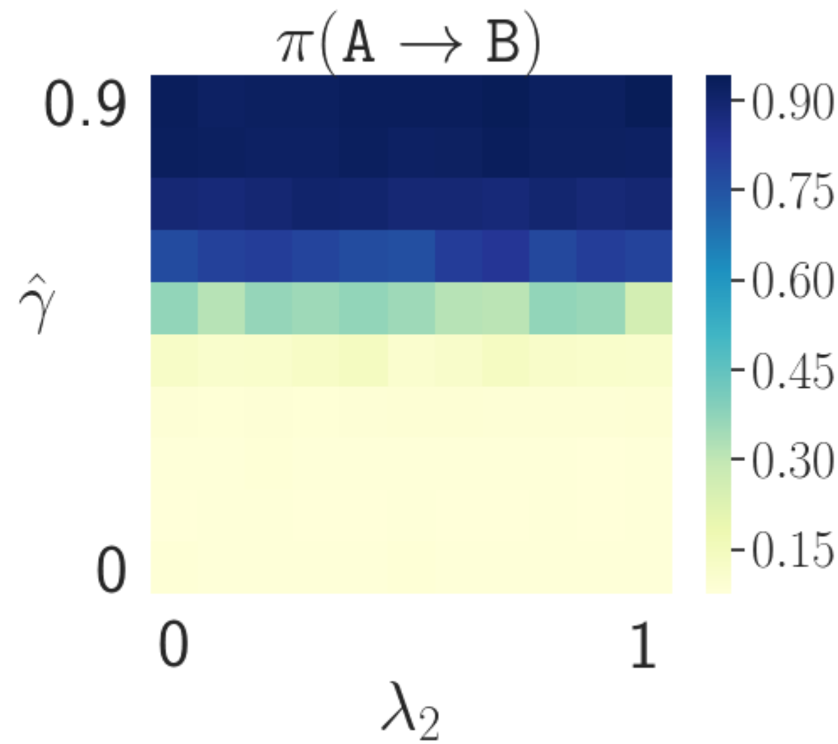
$$F_t^{(2)} \doteq c(S_{t-1}) \rho_{t-1} \nabla \log \pi(S_{t-1}, A_{t-1}) + \hat{\gamma} \rho_{t-1} F_{t-1}^{(1)}$$

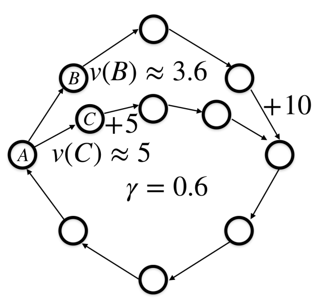
$$Z_t^{(2)} \doteq \hat{\gamma} v_\pi(S_t) F_t^{(2)}$$

$$\lim_{t \to \infty} Z_t^{(1)} + Z_t^{(2)} = \nabla J_{\hat{\gamma}}$$

Geoff-PAC finds the optimal solution in the two circle MDP

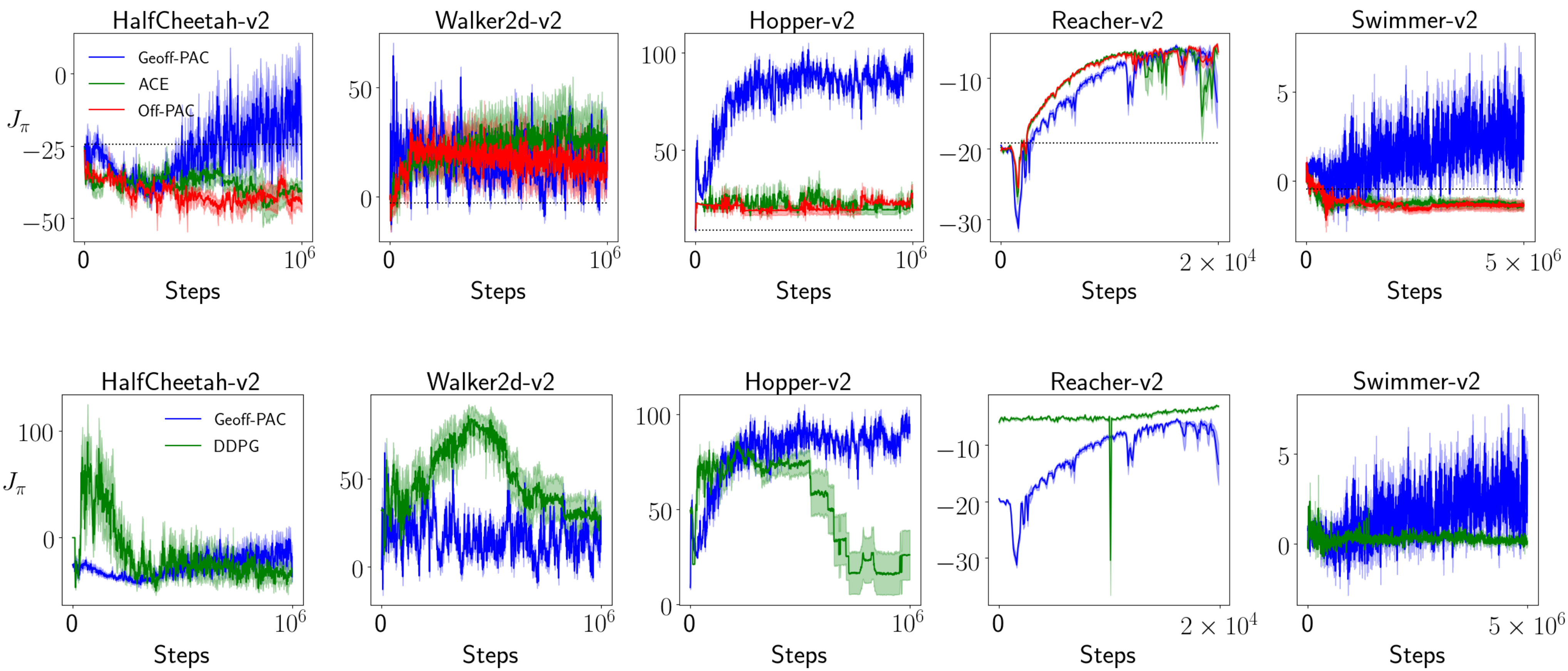






Geoff-PAC scales up to challenging deep RL problems

Evaluation performance of the target policy under a uniformly random behaviour policy



Thanks & Questions

Generalized Off-Policy Actor-Critic

(https://arxiv.org/abs/1903.11329)