Interpolated Policy Gradient[1]

- IPG mixes likelihood ratio gradient with $\widehat{\mathrm{A}}$ which provides unbiased but high-variance gradient estimation, and deterministic gradient through an off-policy fitted criticQ_w

$$\nabla_\theta J(\theta) \approx (1-\nu)\mathbb{E}_{\rho^\pi,\pi}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}(s_t,a_t)] + \nu\mathbb{E}_{\rho^\beta}[\nabla_\theta \bar{Q}_w^\pi(s_t)]$$

- The second term is actually an off-policy action-critic algorithm, but it does not use target policy network but a stochastic policy which enables on-policy exploration

- IPG provides the possibility of **on-policy** Hindsight Experience Replay!

[1]*Gu S, Levine S. 2017. Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning. arXiv preprint arXiv:1706.00387.*

Hindsight Experience Replay[2] for Multi-goal RL

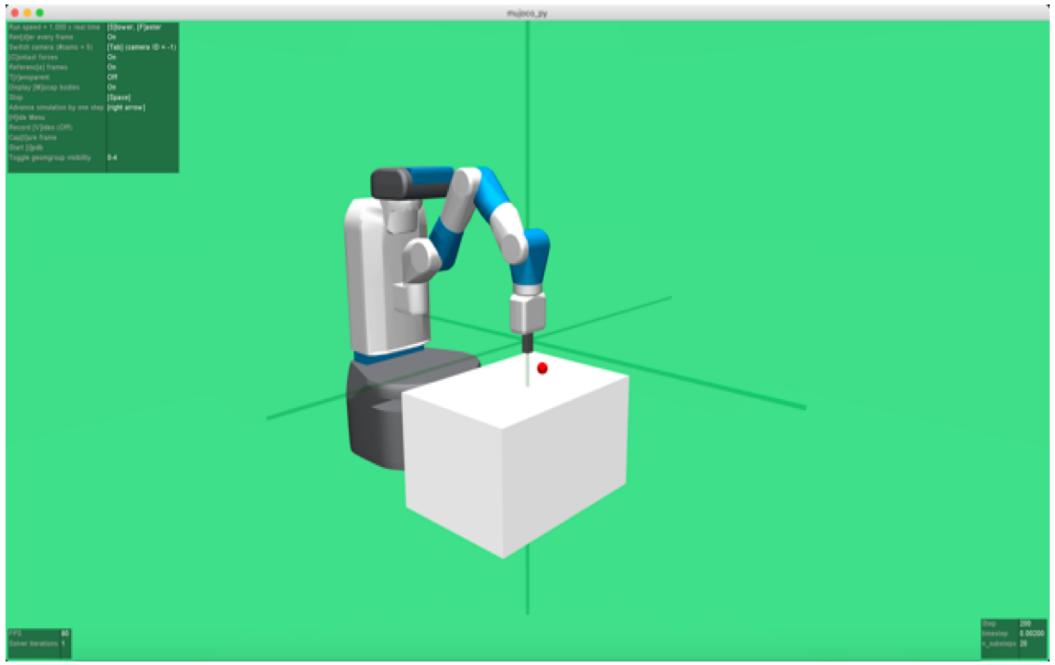
for $t = 0, T-1$ **do**
  $r_t := r(s_t, a_t, g)$
  Store the transition $(s_t||g,\ a_t,\ r_t,\ s_{t+1}||g)$ in R ▷ standard experience replay
  Sample a set of additional goals for replay $G := \mathbb{S}(\textbf{current episode})$
  for $g' \in G$ **do**
    $r' := r(s_t, a_t, g')$
    Store the transition $(s_t||g',\ a_t,\ r',\ s_{t+1}||g')$ in R ▷ HER
  end for
end for

- After experiencing some episode s0, s1, . . . , sT we store in the replay buffer every transition st → st+1 not only with the original **desired goal** used for this episode but also with **future achieved goals**.

k random states which come from the same episode as the transition being replayed and were observed *after* it

[2] *OpenAI, "Hindsight experience replay," in Advances in Neural Information Processing Systems, 2017.*

- Combining HER and IPG generates on-policy HER.
- Hindsight Experience Replay (HER) will speedup the learning process for multi-goal based RL problems.
- IPG doesn’t have a stable performance all the time.



FetchReach-v0 env