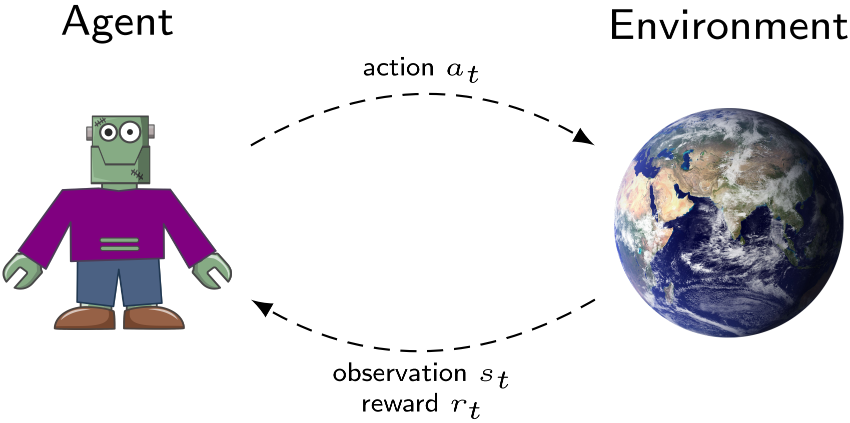
Parameter-based Value Functions

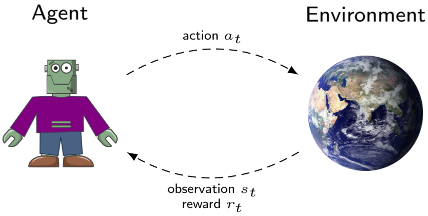
Francesco Faccio (francesco@idsia.ch)
Louis Kirsch and Jürgen Schmidhuber

December 2021


Agent
Environment
action a_t
observation s_t
reward r_t

The RL Framework

- **Markov Decision Process** (Puterman, 2014; Stratonovich, 1960)
  - $\mathcal{S}$ set of states: $s \in \mathcal{S}$
  - $\mathcal{A}$ set of actions: $a \in \mathcal{A}$
  - $\mathcal{P}(s'|s,a)$ markovian transition matrix
  - $R(s,a)$ reward function
  - γ discount factor
  - μ_0 distribution on initial state



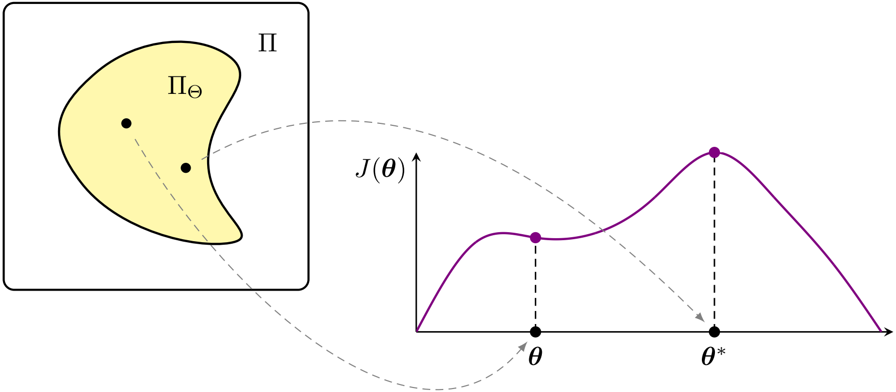
Example

- **Humanoid**:
  - State space: angles and velocities of joints; position of center of mass; momentum
  - Action space: torque on each joint
  - Deterministic state transitions
  - Reward function: $r(s, a) = v_x - 0.005||a||_2^2$, where v_x indicates the forward velocity.

The RL problem

- **RL problem** (Sutton and Barto, 1998): find the optimal policy

$$\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A}) \qquad \boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\Theta} J(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t) | a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t)\right]$$



Traditional Value Functions

(Sutton and Barto, 1998)

- Value functions estimate the return $R_t = \sum_{k=0}^{T-t-1} \gamma^k R(s_{t+k+1}, a_{t+k+1})$ of a policy:
  - State-value function
    $V^{\pi_\theta}(s) := \mathbb{E}_{\pi_\theta}[R_t | s_t = s]$
  - Action-value function
    $Q^{\pi_\theta}(s, a) := \mathbb{E}_{\pi_\theta}[R_t | s_t = s, a_t = a]$
- State and action value functions are related by:

$$V^{\pi_\theta}(s) = \begin{cases} \int_{\mathcal{A}} \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \, \mathrm{d}a & \text{if } \pi_\theta \text{ is stochastic,} \\ Q^{\pi_\theta}(s, \pi_\theta(s)) & \text{if } \pi_\theta \text{ is deterministic.} \end{cases}$$

Traditional Value Functions
(Sutton and Barto, 1998)

- Value functions estimate the return $R_t = \sum_{k=0}^{T-t-1} \gamma^k R(s_{t+k+1}, a_{t+k+1})$ of a policy:
  - State-value function
    $V^{\pi_{\boldsymbol{\theta}}}(s) := \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[R_t | s_t = s]$
  - Action-value function
    $Q^{\pi_{\boldsymbol{\theta}}}(s, a) := \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[R_t | s_t = s, a_t = a]$
- State and action value functions are related by:

$$V^{\pi_{\boldsymbol{\theta}}}(s) = \begin{cases} \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s, a)\, \mathrm{d}a & \text{if } \pi_{\boldsymbol{\theta}} \text{ is stochastic,} \\ Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s)) & \text{if } \pi_{\boldsymbol{\theta}} \text{ is deterministic.} \end{cases}$$

The on-policy policy gradient

- **Problem:** Improve a stochastic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s)\,\mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s.$$

- The gradient is:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s \\
&= \dots \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s, \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s,
\end{aligned}$$

where $d^{\pi_\theta}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \to s', t, \pi_\theta)\,\mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The on-policy policy gradient

- **Problem:** Improve a stochastic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s) \,\mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s.$$

- The gradient is:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s \\
&= \dots \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s, \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s,
\end{aligned}$$

where $d^{\pi_\theta}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \rightarrow s', t, \pi_\theta) \,\mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The on-policy policy gradient

- **Problem:** Improve a stochastic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s) \,\mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s.$$

- The gradient is:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s \\
&= \dots \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s, \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \,\mathrm{d}a \,\mathrm{d}s,
\end{aligned}$$

where $d^{\pi_{\boldsymbol{\theta}}}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \to s', t, \pi_\theta) \,\mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The on-policy policy gradient

- **Problem:** Improve a deterministic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s)\, \mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s))\, \mathrm{d}s.$$

- The gradient is:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s))\, \mathrm{d}s \\ &= \dots \\ &= \int_{\mathcal{S}} d^{\pi_\theta}(s) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_\theta}(s, a)|_{a=\pi_\theta(s)}\, \mathrm{d}s, \end{aligned}$$

where $d^{\pi_\theta}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \to s', t, \pi_\theta)\, \mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The on-policy policy gradient

- **Problem:** Improve a deterministic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s) \, \mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s)) \, \mathrm{d}s.$$

- The gradient is:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s)) \, \mathrm{d}s \\ &= \dots \\ &= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s, a)|_{a=\pi_\theta(s)} \, \mathrm{d}s, \end{aligned}$$

where $d^{\pi_\theta}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \to s', t, \pi_\theta) \, \mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The on-policy policy gradient

- **Problem:** Improve a deterministic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ using data collected from $\pi_{\boldsymbol{\theta}}$.
- Given the objective:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mu_0(s) V^{\pi_{\boldsymbol{\theta}}}(s) \, \mathrm{d}s = \int_{\mathcal{S}} \mu_0(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s)) \, \mathrm{d}s.$$

- The gradient is:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) &= \int_{\mathcal{S}} \mu_0(s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s)) \, \mathrm{d}s \\
&= \dots \\
&= \int_{\mathcal{S}} d^{\pi_{\boldsymbol{\theta}}}(s) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s, a)|_{a=\pi_\theta(s)} \, \mathrm{d}s,
\end{aligned}$$

where $d^{\pi_{\boldsymbol{\theta}}}(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) P(s \to s', t, \pi_\theta) \, \mathrm{d}s$ is the discounted weighting of states encountered starting from $s \sim \mu_0(s)$ and following π_θ.

The off-policy RL problem

- **Problem:** Find the optimal policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ using data collected from a behavioral policy π_b

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Traditional off-policy RL
(Degris et al., 2012; Silver et al., 2014)

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) V^{\pi_{\boldsymbol{\theta}}}(s)\,\mathrm{d}s = \begin{cases} \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is stochastic,} \\ \int_{\mathcal{S}} d^{\pi_b}(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s))\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is deterministic.} \end{cases}$$

where $d^{\pi_b}(s)$ is the stationary distribution of states in the MDP under π_b:

- $d^{\pi_b}(s) = \lim_{t \to \infty} P(s_t = s | s_0, \pi_b)$

The off-policy RL problem

- **Problem:** Find the optimal policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ using data collected from a behavioral policy π_b

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Traditional off-policy RL

(Degris et al., 2012; Silver et al., 2014)

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) V^{\pi_{\boldsymbol{\theta}}}(s)\,\mathrm{d}s = \begin{cases} \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is stochastic,} \\ \int_{\mathcal{S}} d^{\pi_b}(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s))\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is deterministic.} \end{cases}$$

where $d^{\pi_b}(s)$ is the stationary distribution of states in the MDP under π_b:

- $d^{\pi_b}(s) = \lim_{t \to \infty} P(s_t = s | s_0, \pi_b)$

The off-policy RL problem

- **Problem:** Find the optimal policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ using data collected from a behavioral policy π_b

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Traditional off-policy RL

(Degris et al., 2012; Silver et al., 2014)

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) V^{\pi_{\boldsymbol{\theta}}}(s)\,\mathrm{d}s = \begin{cases} \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)\,\mathrm{d}a\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is stochastic,} \\ \int_{\mathcal{S}} d^{\pi_b}(s) Q^{\pi_{\boldsymbol{\theta}}}(s, \pi_{\boldsymbol{\theta}}(s))\,\mathrm{d}s & \text{if } \pi_{\boldsymbol{\theta}} \text{ is deterministic.} \end{cases}$$

where $d^{\pi_b}(s)$ is the stationary distribution of states in the MDP under π_b:

- $d^{\pi_b}(s) = \lim_{t \to \infty} P(s_t = s | s_0, \pi_b)$

The off-policy policy gradient

- When the policy is stochastic (Degris et al., 2012):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\right) \mathrm{d}a \, \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\right) \mathrm{d}a \, \mathrm{d}s.$$

- When the policy is deterministic (Silver et al., 2014):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)} + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s.$$

- **Problems:**
  - In off-policy RL, the gradient of the action value function Q with respect to the policy parameters is often ignored
  - Value functions are defined for a single policy. When value functions are updated to track the learned policy, they forget potentially useful information about old policies

The off-policy policy gradient

- When the policy is stochastic (Degris et al., 2012):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\right) \mathrm{d}a \, \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\right) \mathrm{d}a \, \mathrm{d}s.$$

- When the policy is deterministic (Silver et al., 2014):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)} + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s.$$

- **Problems:**
  - In off-policy RL, the gradient of the action value function Q with respect to the policy parameters is often ignored
  - Value functions are defined for a single policy. When value functions are updated to track the learned policy, they forget potentially useful information about old policies

The off-policy policy gradient

- When the policy is stochastic (Degris et al., 2012):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\right) \mathrm{d}a \, \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_b(a|s) \frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q^{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\right) \mathrm{d}a \, \mathrm{d}s.$$

- When the policy is deterministic (Silver et al., 2014):

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)} + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s$$
$$\approx \int_{\mathcal{S}} d^{\pi_b}(s) \left(\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) \nabla_a Q^{\pi_{\boldsymbol{\theta}}}(s,a)|_{a=\pi_{\boldsymbol{\theta}}(s)}\right) \mathrm{d}s.$$

- **Problems:**
  - In off-policy RL, the gradient of the action value function Q with respect to the policy parameters is often ignored
  - Value functions are defined for a single policy. When value functions are updated to track the learned policy, they forget potentially useful information about old policies

PVFs

Parameter-based Value Functions

(Faccio et al., 2021)

- Parameter-based State-Value Function (**PSVF**)

  $V(s, \boldsymbol{\theta}) := \mathbb{E}[R_t | s_t = s, \boldsymbol{\theta}]$

- Parameter-based Action-Value Function (**PAVF**)

  $Q(s, a, \boldsymbol{\theta}) := \mathbb{E}[R_t | s_t = s, a_t = a, \boldsymbol{\theta}]$

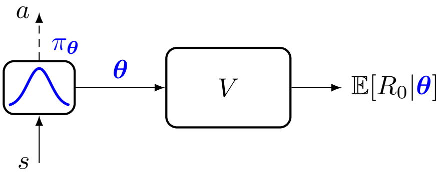
- Parameter-based Start-State-Value Function (**PSSVF**)

  $V(\boldsymbol{\theta}) := \mathbb{E}_{s \sim \mu_0(s)}[V(s, \boldsymbol{\theta})]$

Parameter-based Start-State Value Function

- Stochastic or deterministic policies
- Find the policy $\pi_{\boldsymbol{\theta}}$ maximizing $J(\boldsymbol{\theta})$:

$$J(\boldsymbol{\theta}) = \mathbb{E}[R_0|\boldsymbol{\theta}] = V(\boldsymbol{\theta})$$



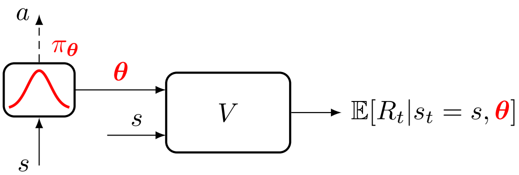
- Taking the gradient of $J(\boldsymbol{\theta})$ we obtain:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$$

Parameter-based State-Value Function

- Stochastic or deterministic policies
- Find the policy $\pi_{\boldsymbol{\theta}}$ maximizing $J(\boldsymbol{\theta})$:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) V(s, \boldsymbol{\theta}) \, \mathrm{d}s$$



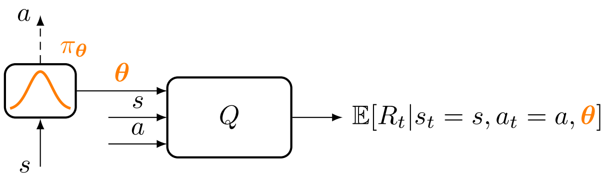
- Taking the gradient of $J(\boldsymbol{\theta})$ we obtain:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{s \sim d^{\pi_b}(s)} [\nabla_{\boldsymbol{\theta}} V(s, \boldsymbol{\theta})]$$

Parameter-based Action-Value Function

- Stochastic policies
- Find the policy $\pi_{\boldsymbol{\theta}}$ maximizing $J(\boldsymbol{\theta})$:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) Q(s, a, \boldsymbol{\theta}) \, \mathrm{d}a \, \mathrm{d}s$$



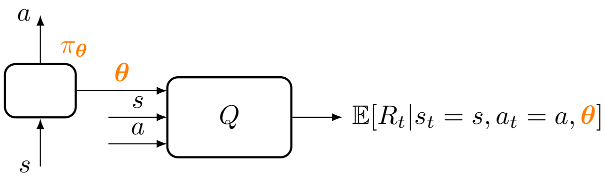
- Taking the gradient of $J(\boldsymbol{\theta})$ we obtain:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{s\sim d^{\pi_b}(s), a\sim\pi_b(.|s)} \left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_b(a|s)} \left(Q(s, a, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) + \nabla_{\boldsymbol{\theta}} Q(s, a, \boldsymbol{\theta}) \right) \right]$$

Parameter-based Action-Value Function

- Deterministic policies
- Find the policy $\pi_{\boldsymbol{\theta}}$ maximizing $J(\boldsymbol{\theta})$:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi_b}(s) Q(s, \pi_{\boldsymbol{\theta}}(s), \boldsymbol{\theta}) \, \mathrm{d}s$$

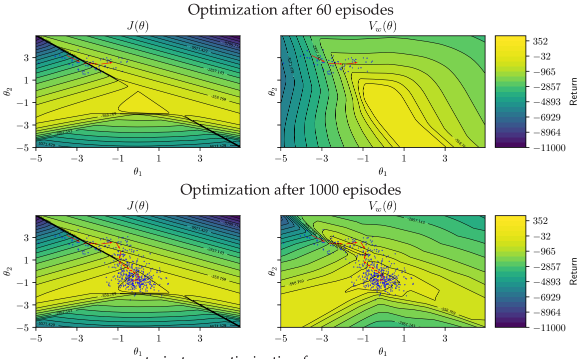


- Taking the gradient of $J(\boldsymbol{\theta})$ we obtain:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{s \sim d^{\pi_b}(s)} \left[\nabla_a Q(s, a, \boldsymbol{\theta})|_{a=\pi_{\boldsymbol{\theta}}(s)} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(s) + \nabla_{\boldsymbol{\theta}} Q(s, a, \boldsymbol{\theta})|_{a=\pi_{\boldsymbol{\theta}}(s)} \right]$$

Actor-Critic algorithm

- PSSVF on LQR using deterministic shallow policies



trajectory optimization for π_θ: → π_b: •

Off-policy actor-critic with PVFs

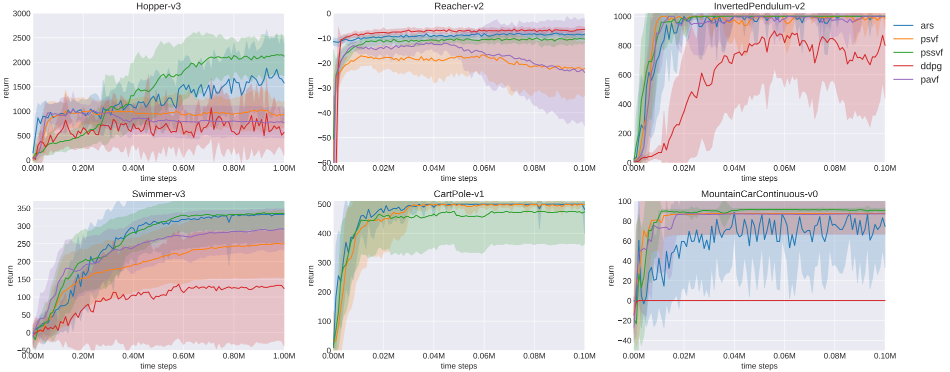
Given the behavioral π_b, find $\pi_{\boldsymbol{\theta}}$ **maximizing** $J(\boldsymbol{\theta})$:

1. Collect data with π_b (expensive in RL)
2. Use data to train $V(\boldsymbol{\theta})$, $V(s, \boldsymbol{\theta})$ or $Q(s, a, \boldsymbol{\theta})$
3. Find $\pi_{\boldsymbol{\theta}}$ following $\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})$ (offline optimization)
4. Set new behavioral $\pi_{\boldsymbol{\theta}} \leftarrow \pi_b$
5. Repeat until convergence

Experiments

Shallow policies

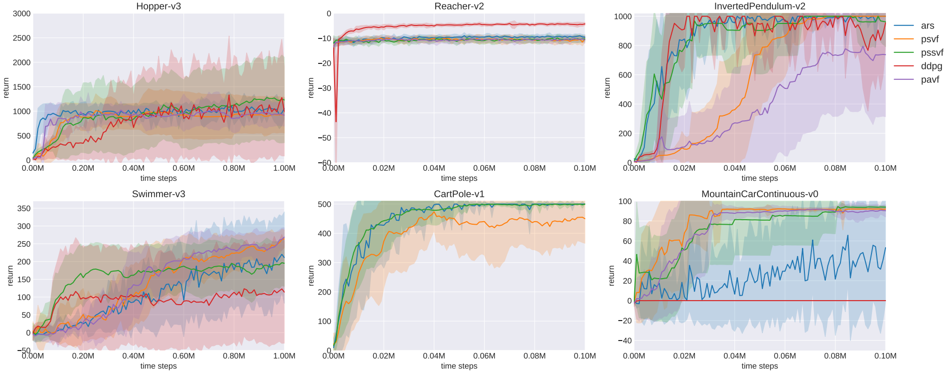
- Comparison with DDPG (Lillicrap et al., 2015) and ARS (Mania et al., 2018)



Experiments

Deep policies

- Comparison with DDPG (Lillicrap et al., 2015) and ARS (Mania et al., 2018)



Discussion and Conclusions

- Contributions
  - A new class of value functions that generalize across policies
  - Novel off-policy policy gradient theorems
  - New off-policy actor-critic algorithms
  - Experimental results comparable with state-of-the-art algorithms
- Future works
  - Parameter generators
  - Policy embedding - dimensionality reduction
  - Convergence results
  - Extension to RNNs

Discussion and Conclusions

- Contributions
  - A new class of value functions that generalize across policies
  - Novel off-policy policy gradient theorems
  - New off-policy actor-critic algorithms
  - Experimental results comparable with state-of-the-art algorithms
- Future works
  - Parameter generators
  - Policy embedding - dimensionality reduction
  - Convergence results
  - Extension to RNNs

Thank You for Your Attention!

References

Degris, T., White, M., and Sutton, R. S. (2012). Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*.

Faccio, F., Kirsch, L., and Schmidhuber, J. (2021). Parameter-based value functions. *arXiv preprint arXiv:2006.09226*.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Mania, H., Guy, A., and Recht, B. (2018). Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1800–1809.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *ICML*.

Stratonovich, R. (1960). Conditional Markov processes. *Theory of Probability And Its Applications*, 5(2):156–178.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.