HSLU Lucerne University of Applied Sciences and Arts

Policy Gradient Methods Summary

go right

Wait, I am still calculating Q-values.....

Reinforcement Learning
December 22, 2022

Policy Gradient Methods

- Learn the policy function directly

$$\pi(a|s, \boldsymbol{\theta}) = \Pr\{A_t = a | S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}$$

- Constraints:

$$\pi(a|s, \boldsymbol{\theta}) > 0, \quad \text{for all } a \in \mathcal{A}, s \in \mathcal{S}$$

$$\sum_a \pi(a|s, \boldsymbol{\theta}) = 1, \quad \text{for all } s \in \mathcal{S}$$

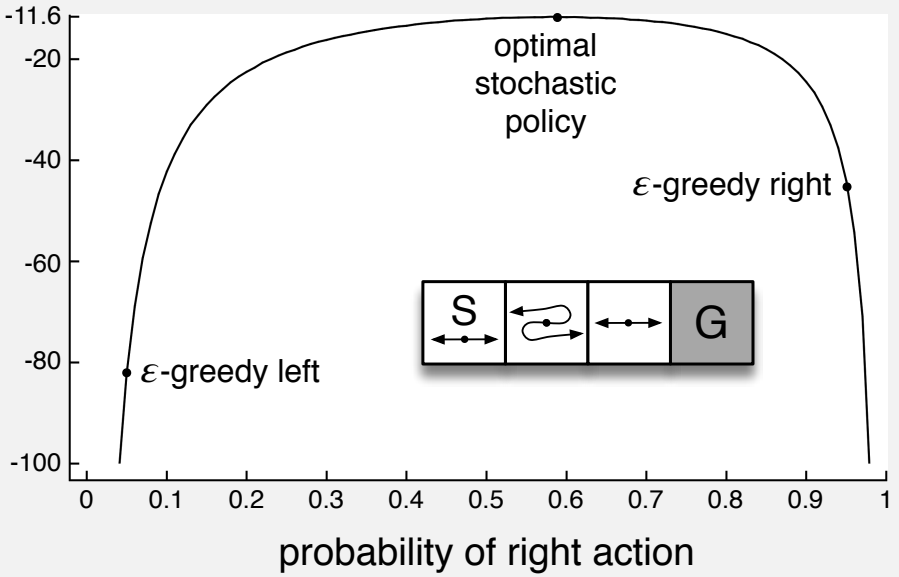
  which can be enforced by using the softmax function

Advantages of Policy Parametrization

A parametrized policy

- can approach a deterministic policy over time (in comparison to epsilon-greedy, which will always explore)
- can model stochastic policies

$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(\mathrm{S})$



Policy Gradient Theorem

- Goal: optimize the total return from a (particular) state

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

- The gradient of the Loss function is proportional to the gradient of the policy

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta})$$

REINFORCE: Monte Carlo Policy Gradient

- Update the weights according to:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|s, \boldsymbol{\theta})} \\ &= \boldsymbol{\theta}_t + \alpha G_t \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})\end{aligned}$$

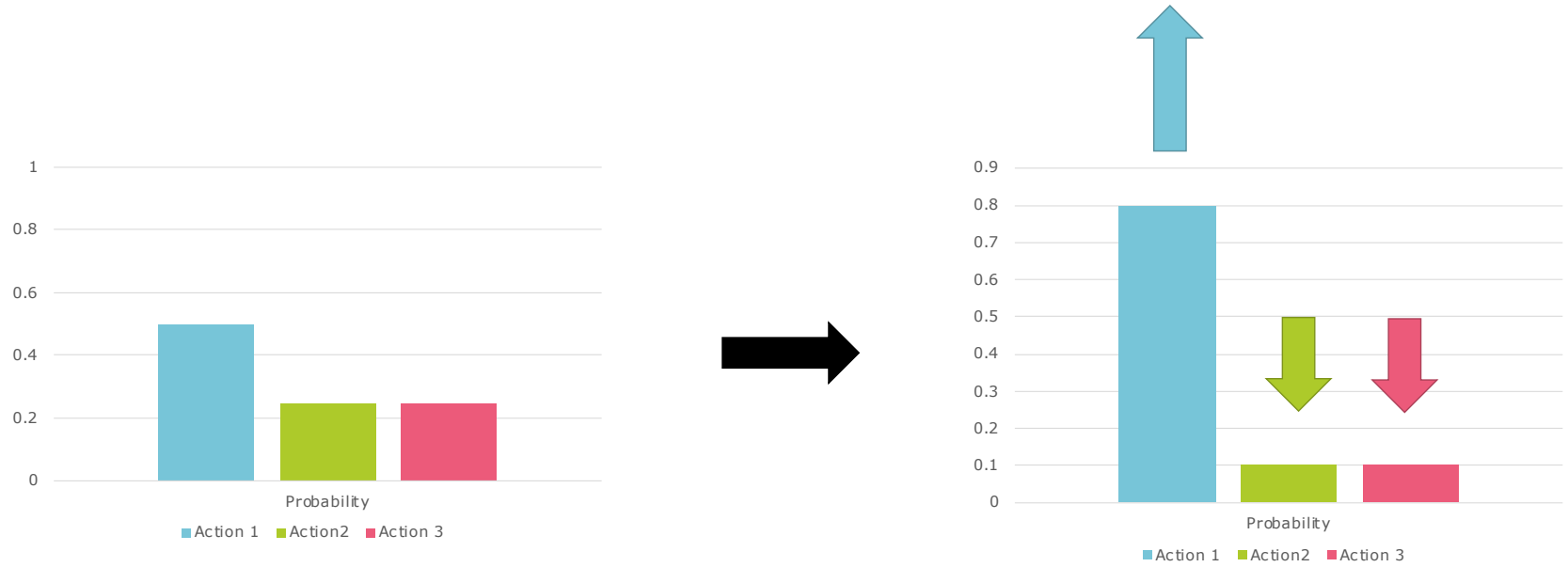
- The second line can be derived from

$$\nabla \ln(f(x)) = \frac{\nabla f(x)}{f(x)}$$

- This is gradient **ascent**, as we want to maximize the return

How does Policy Gradient work?

- The gradient will push the selected action to have a higher probability (at the expense of the others)
- The amount that it will be pushed depends on the return G



Actor Critic: Policy Gradient Method with Critic

- For the 1-step return target function, we need a **value** function

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha\left(R_{t+1} + \gamma\hat{v}(S_{t+1}, \mathbf{w}) - \hat{v}(S_t, \mathbf{w})\right)\nabla \ln \pi(A_t|S_t, \boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t + \alpha\,\delta_t\,\nabla \ln \pi(A_t|S_t, \boldsymbol{\theta}_t)\end{aligned}$$

- The value function is approximated too

One step Actor-Critic (episodic) for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input:
  a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
  a differentiable state-value function parametrization $\hat{v}(s, \mathbf{w})$
  step sizes $\alpha^{\boldsymbol{\theta}} > 0, \alpha^{\mathbf{w}} > 0$
Initialize:
  policy parameters $\boldsymbol{\theta}$ and state-value weights $\mathbf{w}$

Loop for each episode:
  Initialize S, first state of episode
  $I \leftarrow 1$
  For each time step of the episode:
    Choose $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S, \mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

TRPO and PPO Methods

In policy gradient methods, the update is calculated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \nabla J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}$$

The new values for $\boldsymbol{\theta}$ should be near to the old values, as we use a small step size, however, the **policy** could still change significantly with a change of $\boldsymbol{\theta}$

TRPO and PPO Methods ensure that the policy does not change too much

PPO

$$p_t(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)}$$

$$L^{CLIP}(\boldsymbol{\theta}) = \mathbb{E}_t[\min(r_t(\boldsymbol{\theta})A_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)A_t]$$

- Clip the advantage function when it is outside of the comfort interval bounded by $1 \pm \epsilon$
- Due to the minimum operation, the probability ratio is ignored outside the zone only if it would improve the objective, it is included if it makes the objective worse