# Reinforcement Learning: Tutorial 11

# Policy gradient methods: PGT, DPG & evaluation

Week 6
University of Amsterdam

Milena Kapralova
October 2024

# Check-in

- How is it going?
- How is HW5?
- If you have any feedback so far, please mail me at *m.kapralova@uva.nl*

# Outline

1. Admin
2. Policy gradient methods exercises
3. Ask anything about HW5

# Admin

- Please direct any questions about grading to Pieter Pierrot
- Any questions?

# Tutorial 11 Overview

1. Policy gradient methods exercises
2. Ask anything about HW5

# Tutorial 11 Overview

1. Policy gradient methods exercises
   - Questions 10.1-10.2
2. Ask anything about HW5

# Q 10.1 *Exam Question: Actor-critic algorithm

**One-step Actor–Critic (episodic), for estimating $\pi_\theta \approx \pi_*$**

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize $S$ (first state of episode)
  $I \leftarrow 1$
  Loop while $S$ is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S', R$
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if $S'$ is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

1. In the lecture, we have used a slightly different actor-critic update, namely:

$$\theta_{t+1} = \theta_t + \alpha\, \hat{q}_{\mathbf{w}}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

There are two main differences between that update and the update in the algorithm in the figure above (other than the discount factor). Describe the two differences, and for each difference, say why the authors might have chosen to perform the update this way.

# Q 10.1 *Exam Question: Actor-critic algorithm

**One-step Actor–Critic (episodic), for estimating $\pi_\theta \approx \pi_*$**

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize $S$ (first state of episode)
  $I \leftarrow 1$
  Loop while $S$ is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S', R$
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if $S'$ is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

a) Instead of $q(s,a)$, $\delta$ is used. This relies on $v$ instead of $q$. The reason can be that $v$ is easier to learn than $q$ (as it is a function of just one variable).
b) $\delta$ also includes a baseline. The baseline reduces variance just as in REINFORCE.

# Q 10.1 *Exam Question: Actor-critic algorithm

**One-step Actor–Critic (episodic), for estimating $\pi_\theta \approx \pi_*$**

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize $S$ (first state of episode)
  $I \leftarrow 1$
  Loop while $S$ is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S', R$
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if $S'$ is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

2. The algorithm in figure above uses a discount factor $\gamma$, which we have not considered in the lecture. There is a special case where we can ignore the factor $I$ in the policy update and keep overall convergence behavior (even if $\gamma < 1$). What is this case? *Hint: For a given state, the inclusion of the factor I does not change the direction of the expected gradient* $\mathbb{E}_{a\sim\pi}[I\nabla \log \pi(A|S,\theta)]$.

# Q 10.1 *Exam Question: Actor-critic algorithm

2. The algorithm in figure above uses a discount factor $\gamma$, which we have not considered in the lecture. There is a special case where we can ignore the factor $I$ in the policy update and keep overall convergence behavior (even if $\gamma < 1$). What is this case? *Hint: For a given state, the inclusion of the factor $I$ does not change the direction of the expected gradient* $\mathbb{E}_{a\sim\pi}[I\nabla \log \pi(A|S,\theta)]$.

If the policy is tabular. The only thing that changes with the discounting is how much the error in one state contributes compared to the error in other states $\rightarrow$ if changing a parameter changes the value in multiple states, their relative weight will change if a state is usually encountered early or late in trajectories.
With tabular setting there is no such trade-off. The extra $I$ term will sometimes make learning rates smaller, but that doesn't change the average direction the policy is updated in, so it will not change the local maxima in $J(\theta)$ that the policy can converge to.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

1. $$\mathbb{E}_\tau \left[ \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

1. $$\mathbb{E}_\tau \left[ \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

   Unbiased, note that this is the normal REINFORCE update.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

2. $$\mathbb{E}_\tau \left[ \sum_{t'=1}^{T} r_{t'} \sum_{t=t'}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

2. $$\mathbb{E}_\tau \left[ \sum_{t'=1}^{T} r_{t'} \sum_{t=t'}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

   These are the terms we leave out in REINFORCE v2 (the $a$'s after the corresponding $r$), which are expectation 0. So we only look at the noise and we leave out the signal. Thus, it is biased (and completely useless) as estimator for the policy gradient.
   Note: In Ex. 9.2.2 (last week's tutorial), each gradient term was weighted by the sum of all rewards up to the current time step. Equivalently here, each reward is weighted by the sum of all gradient terms after the current time step.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

3

$$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left( G_t(\tau) - \hat{V}_w(s_t) \right) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

3

$$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left( G_t(\tau) - \hat{V}_w(s_t) \right) \right]$$

Unbiased, this is REINFORCE with baseline. The expected value of the subtracted term is zero (RL:AI, p.329 & Deisenroth et al, p.27):

$$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left( G_t(\tau) - \hat{V}_w(s_t) \right) \right]$$

$$= \mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) G_t(\tau) \right] - \mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \hat{V}_w(s_t) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

$$\text{Then} \quad \mathbb{E}_{\tau\sim\pi_\theta} \sum_{t=1}^{T} \Big[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{V}_w(s_t)\Big]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{s_{0:t},a_{0:t-1}} \Big[\mathbb{E}_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{V}_w(s_t)\ ]\Big]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{s_{0:t},a_{0:t-1}} \Big[\hat{V}_w(s_t) \cdot \mathbb{E}_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)\ ]\Big]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{s_{0:t},a_{0:t-1}} \Big[\hat{V}_w(s_t) \cdot \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\ ]\Big]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{s_{0:t},a_{0:t-1}} \Big[\hat{V}_w(s_t) \cdot 0\Big] = 0$$

Based on *this blogpost*. In the last step we used the log derivative trick.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

4. $$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) q_\pi(a_t, s_t) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

4. $$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) q_\pi(a_t, s_t) \right]$$

   Unbiased (return is true $q_\pi$, which gives the correct target for the policy update).

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

5. $$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left( G_t(\tau) - \hat{q}_w(s_t, a_t) \right) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

5. $$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left( G_t(\tau) - \hat{q}_w(s_t, a_t) \right) \right]$$

   Biased, baseline depends on $a_t$ which depends on trajectory.

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

6

$$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{q}_w(s_t, a_t) \right]$$

# Q 10.2 *Exam Question: Policy gradient methods

For each of the following equations, determine whether they are unbiased estimators of the gradient $\nabla_\theta \mathbb{E}_\tau [G(\tau)]$ of undiscounted returns $G(\tau)$. Briefly explain your answer.

6

$$\mathbb{E}_\tau \left[ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{q}_w(s_t, a_t) \right]$$

Biased since $q_w$ is an approximation to $q_\pi$ and might not be exact. (In an optional module we covered the compatible function approximation theorem which tells us that the estimator is only unbiased if the value function is compatible s.t. expected value remains unchanged.)


WELL DONE!

# Tutorial 11 Overview

1. Policy gradient methods exercises
2. Ask anything about HW5
   - Questions 9.4, 10.3-10.4

# Ask anything about HW4

- 9.4: Theory
- 10.3: Theory
- 10.4: Coding (+ Little bit of theory)

# That's it!

See you tomorrow